A mechanistic model predicting cell surface presentation of peptides by MHC class I proteins, considering peptide competition, viral intracellular kinetics and host genotype factors

Ruth Charlotte Eccleston

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London.

Centre of Maths and Physics in Life-sciences and Experimental biology and the

Department of Chemistry

University College London

September 10, 2017

I, Ruth Charlotte Eccleston, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

Major histocompatability complex class I (MHC-I) proteins present short fragments of pathogenic or cancerous proteins (peptides) on the surface of infected cells for recognition by T lymphocytes which are stimulated upon recognition of foreign peptides. Due to the diversity of peptide sequences and the sequence-specificity of MHC-I alleles, being able to determine which peptides will be presented by which MHC-I alleles and in what proportion could be important for the development of vaccines and treatments based on the presented peptiodome. Machine learning tools, trained on experimental data, are widely used to predict immunogenic peptides. However they are unable to account for the impact the intracellular kinetics of the pathogenic or cancerous protein which will greatly influence the resultant peptidome. Here we describe a mechanistic model of peptide presentation, validated against experimental data, which accounts for intracellular peptide concentration, and can predict the relative cell surface presentation of competing peptides with varying affinities for MHC-I proteins. We demonstrate how combining this mechanistic model with the intracellular kinetics of HIV proteins can provide insight in to the experimentally reported immunogenicity of the viral protein Gag, and show how such a model can be used to predict the most abundant viral peptides presented on the cell surface. Similarly, we predict the HeLa cell peptidome and demonstrate how a simple metric can be used to approximate the abundance of a peptide based solely on protein synthesis and degradation, peptide-MHC affinity and proteasomal cleavage.

Acknowledgements

The past four years have been as mind-blowingly challenging as they have been rewarding, with many months spent banging my head against a brick wall for the smallest of victories. I would like to thank my supervisors Neil and Peter for all their invaluable help, guidance and support, I have enjoyed working with you both. I would also like to thank our collaborators at Southampton, Tim and Denise, without whom Chapter 3 of this thesis would not have been possible.

I am eternally grateful to UCL CoMPLEX for offering me a place on their DTC, and for all they have done over the past four years to support me and all my cohort during our MRes and PhD. To my CoMPLEX friends, especially Lourdes, you have made this experience so much fun and I am so glad I met you. I will fondly remember all our times at Cumberland Lodge and especially the trip to Plymouth during our MRes year.

To *Dr* Laura Parshotam, I could not have got through the past three years - both PhD and life in general - without you. You are a wonderful human being and I am so lucky to know you.

I would also like to thank the rest of the CCS, past and present, for putting up with my not-so-conducive-to-work office antics.

To Saavidra Perera thank you for getting me through my MSci in the first place and for allowing me to vent over countless long phone calls - you are the wittiest girl in Physics after all!

To my mum thank you for always believing in me, for supporting me and knowing me better than I know myself by encouraging me to do science in the first place!

Finally, I would like to thank Barnaby Walker for his patience and love, and for brightening every day with his unique brand of humour. I wouldn't be here without you.

Contents

List of Figures

List of Tables

List of Algorithms

Author Publications

Journal Articles

- RC Eccleston, S Wan, N Dalchau and PV Coveney. (2017) **The role of multi-scale protein dynamics in antigen presentation and T lymphocyte recognition**. *Frontiers in Immunology*, 797(8).

Manuscripts Under Consideration

- RC Eccleston, PV Coveney and N Dalchau. **Host genotype and time dependent antigen presentation of viral peptides: predictions from theory**.

- RC Eccleston, D Boulanger, A Phillips, PV Coveney, TJ Elliot and N Dalchau. **A mechanistic model for predicting cell surface presentation of competing peptides by MHC class I molecules**.

Chapter 1

Introduction

1.1 The Immune System

The immune system is a complex network of organs, cells and proteins working together to protect the body from infection by pathogens, such as viruses and bacteria, as well as seeking out and destroying the body's own cells that may have become cancerous due to mutations. In general, the immune system is split in to two: the adaptive immune system and the innate immune system. Although these two systems are largely considered separately, there is evidence of an interface between them[1, 2, 3]. In this thesis, however we will be focussing solely on the adaptive immune system.

During initial infection by a pathogen the first line of defence is the innate immune system. The responses of the innate immune system are not pathogen specific, but instead recognise general conserved pathogen features, known as *pathogen-associated immunostimulants*. The recognition of these molecules results in inflammatory responses and phagocytosis, where cells such as neutrophils and macrophages "eat" an infected cell[4].

In contrast, the response of the adaptive immune system is pathogen specific and can provide long lasting immunity to the pathogen in question. Something that triggers a response from the adaptive immune system is known as an *anti*body *gen*erator or *antigen*, usually in the form of a pathogenic protein. There are two classes of adaptive immune response once an antigen has been detected - antibody responses and cell mediated immune responses - and they both involve white blood cells called *lymphocytes*. Lymphocytes

known as B-cells, which are produced in the bone marrow, are responsible for antibody responses, whereas another type of lymphocyte known as T-cells or cytotoxic T lymphocytes (CTLs), are produced in the thymus and are responsible for cell mediated responses.

When B-cells are activated they produce antibodies known as *immunoglobulins* which bind to the antigen responsible for activating the B-cell and inactivate the pathogen. T-cells are activated by foreign antigens presented on the surface of cells, and once activated they work to directly destroy the infected cell.

In this thesis we will focus on the cell mediated immune response of the adaptive immune system, specifically on the intracellular or endogenous antigen processing pathway.

1.2 The Antigen Processing Pathway

When a human cell is infected by viruses or bacteria, the adaptive immune system is alerted to its presence by the presentation of short lengths of pathogenic proteins, known as peptides, on the cell surface. These peptides are presented by proteins encoded by the Major Histocompatability Complex (MHC) region of the genome, which encodes for three classes of genes that make up part of the innate immune system: MHC Class I, Class II, and Class III molecules. The presentation of peptides by these MHC molecules alerts T-cells to the presence of the infection. The immune system will focus its response on only a few peptides out of the many possible sequences in a process known as *immunodominance*[5]. Immunodominance is when the immune response is directed at only a few of the possible antigenic peptides. Peptides that are recognised by the immune system are known as epitopes, and more specifically, peptides recognised by T-cells are known as T-cell epitopes. The immunodominance of the set of epitopes and the efficiency of the T-cell (thymus cell) response are determined by the cell surface abundance of the MHC-I-peptide complex[6, 7], the affinity of the complex with the T-cell Receptor (TCR), and the frequency of the T-cell precursor[8].

1.2.1 The T-cell receptor

The TCR is a complex of proteins found on the surface of T-cells that binds with the pMHC complex on the surface of the cell. Stimulation of the TCR results in a signalling cascade that activates the T-cell response against the cell infected with the pathogen. The

co-receptors CD4 and CD8 enhance the T-cell signal by simultaneously binding to surface peptide bound MHC-II and MHC-I respectively. TCRs are generally specific to the peptide-MHC complex as they are produced by gene rearrangement during T-cell development in the thymus, meaning each TCR receptor can be very different from one another, and therefore each person produces a very large number of TCRs. This results in TCRs that are very sensitive to only a small number of pMHC complexes out of the many thousands of possibilities, and are highly specific to certain pMHC combinations.

The nature of T-cell antigen recognition is subject to much debate. Wu et al 2002[9] provide evidence that the initial docking between the pMHC complex and the TCR is dictated by the contacts of the MHC, whilst the peptide contacts stabilize the binding. However Burrows et al. 2010[10] report that the TCR is able to maintain peptide recognition following mutations in three important MHC contacts, suggesting the peptide contacts are more important than the MHC.

When a TCR binds with pMHC a structure termed the *immunological synapse* (IS) is formed. Kupfer and collegues first viewed this structure as two concentric rings of molecules[11], that were named the central supramolecular activation cluster (cSMAC) and peripheral supramolecular activation cluster (pSMAC). These bullseye shaped synapses are observed between CD4+ T-cells and B-cell lymphoma tumour cells, and when CD8+ T-cells interact with target cells. It was initially thought that this bullseye formation was crucial for sustained TCR signalling. However, this bullseye structure is not universal to immunological synapses, as T-cells contacting dendritic cells (DCs) instead have a multifocal IS, and T-cell activation is possible before the fully mature IS is formed. The formation of this synapse initiates down-stream signalling, including calcium release, actin remodelling and finally T-cell activation.

A single pMHC complex is able to stimulate the release of intracellular calcium within the T-cell[12]; however, several hundreds of complexes are required for full T-cell activation and Lavoie et al. [13] report a quantitative relationship between pMHC concentration and the magnitude of T-cell activation. T-cells are activated in peripheral lymphoid organs when they recognise a cell infected with a pathogen, where they then proliferate and differentiate in to effector cells, which can then kill the cells infected with

that specific pathogen.

1.2.2 Peptide-MHC binding

Peptides can be presented by both MHC class I and class II proteins. Both MHC Class I and II binding grooves are structurally similar, made up of two α helices and eight β strands. The MHC class I binding groove, however, is closed, whereas the MHC class II binding groove is open. This means MHC class I proteins can only form stable complexes with peptides of length 8-13 amino acids long, although nonamers are most preferred as this is the length of the binding groove, whereas MHC class II can bind peptides between 14-20 amino acids long. Different MHC proteins will preferentially bind different peptide sequences, due to the polymorphism of the amino acids that make up the peptide-binding groove. Both MHC class I and II present both pathogenic and 'self' peptides (derived from degradation of the own host cell protein). Autoimmune diseases result from T-cells recognizing self peptides and mounting a T-cell response against them. The innate immune system discriminates self from non-self by positive selection for TCRs with a low affinity for self-peptide-MHC complexes and negative selection for TCRs with high affinity for self-peptide-MHC complexes[14].

Human Leukyocyte Antigens (HLA) are the human MHC proteins. The MHC Class I loci are split in to three genes HLA-A, HLA-B and HLA-C, all of which are highly polymorphic and so are associated with many different alleles, differing by one or more amino acid substitution. There are six genes in the MHC Class II loci: HLA-DPA1, HLA-DPB1, HLA-DQA1, HLA-DQB1, HLA-DRA and HLA-DRB1. Each individual can express up to two different alleles from each of the three HLA loci, inherited from their parents, and the HLA loci are the most polymorphic in the human genome, with HLA-B alone having more than 400 alleles, whilst in entirety, the Class I and II loci are associated with over 1300 possible alleles.

1.2.3 The exogenous and endogenous pathways

In general, peptides which are produced via intracellular processing of pathogenic proteins bind to MHC class I protein, as part of what is known as the endogenous antigen processing pathway. Peptides in complex with MHC class I interact with the CD8 recep-

tors of T-cells known as cytotoxic T-lymphocytes.

In the exogenous or extracellular pathway, extracellular antigens are taken in to the cell via intracellular vesicles known as endosomes. As the endosomes progress further in to the cell, the antigen is broken down in to peptides and made available for binding to MHC Class II molecules for cell surface presentation. Peptides in complex with MHC class II molecules interact with the CD4 receptors of T-cells known as T-helper (Th) cells. MHC Class II molecules are expressed in professional antigen presenting cells (APC) such as macrophages, B cells and dendritic cells. When CD4+ T-cells are activated they produce a range of cytokines and chemokines. One type of CD4+ T-helper cell known as Th1 is known to produce the cytokine inferon gamma (IFN-γ), which increases intracellular production of MHC and enhance the cytotoxic function of CD8+ T-cells.

In this work we will focus on mathematically modelling the endogenous or intracellular antigen processing pathway (Figure 1.1), by approximately solving a system of coupled differential equations to simulate the concentration of each species present in the system (see Chapter 2 for more details). In the endogenous pathway, peptides are produced when ubiquitin tagged cytoplasmic proteins are degraded by the proteasome. Peptides not only derive from the degradation of mature proteins (retirees), but also from other sources such as Defective Ribosomal Products (DRiPs)[15, 16]. DRiPs are defined as "prematurely terminated polypeptides and misfolded polypeptides produced from translation of *bona fide* mRNAs in the proper reading frame"[15], and are thought to be a significant source of peptide. In this work however, we will not consider the impact of DRiPs on antigen presentation as the exact nature of DRiPs remains highly controversial, and the data we require to include them in the models presented in this thesis are so far not available. We include a section in Chapter 6 discussing how future work could include DRiPs.

Peptides are transported to the endoplasmic reticulum (ER), via the transporter associated with antigen processing (TAP), where they are available for binding to MHC molecules. Peptides compete for binding to the peptide loading complex (PLC), which is made up of TAP, along with the chaperon molecules tapasin, calreticulin and ERp57[17]. Tapasin acts as a filtering mechanism by increasing the unbinding rate of the peptide from

the MHC-I molecule, and so only peptides with a high affinity for the MHC in question stay bound long enough to be presented on the cell surface, thus influencing peptide immunodominance[18]. Recently, another chaperone molecule TAPBPR (a tapasin homologue), has also been found to act like tapasin to enhance peptide optimization[19, 20]. The unbinding rate of the peptide from the MHC complex is therefore an important factor in determining the cell surface abundance, and thus the immunogenicity of the peptide.

Another prerequisite for presentation is a high abundance of the peptide in the ER, as this will increase the likelihood of successful MHC-I binding when competing against thousands of other self and pathogenic peptides. The abundance of a peptide in the ER is determined by the synthesis and degradation of the protein from which the sequence is cleaved, the probability of cleavage of that sequence by the proteasome, and the affinity of that peptide with TAP.

Furthermore, the timing of the appearance of viral peptides on the cell surface following infection, and thus the intracellular kinetics of the viral proteins, also influences the T-cell response[21]. When a CD8 T-cell recognises a foreign antigen it is activated in the draining lymph nodes and releases cytotoxic granules to induce apoptosis and kill the infected cell. CD8 T-cells also induce the production of IFN-γ. Some cross-presentation can occur between the exogenous and endogenous pathways, such as presentation of peptides taken up from the extracellular environment being presented by MHC-I.

1.2.4 T-cell epitopes

The development of T-cell vaccines for viruses such as HIV, or immunotherapy for diseases such as cancer, requires mapping of the peptide hierarchies and identification of immunodominant peptides, known as epitopes, and so epitope discovery is an important area of research. For example, Livingston et al.[23] successfully immunized mice with a pool of four HLA-DR-restricted HIV Th cell epitopes. T-cell vaccines will not prevent infection, but aim to allow the immune system to control infection and limit or completely stop the development of disease caused by the infection.

A correlation between the ability to contain HIV infection and the presence of strong HIV-specific CD8+ T-cells and Th cell responses has been observed[24]. Vaccines stimulating CD4+ or CD8+ T-cell responses to HIV may therefore act to delay or prevent

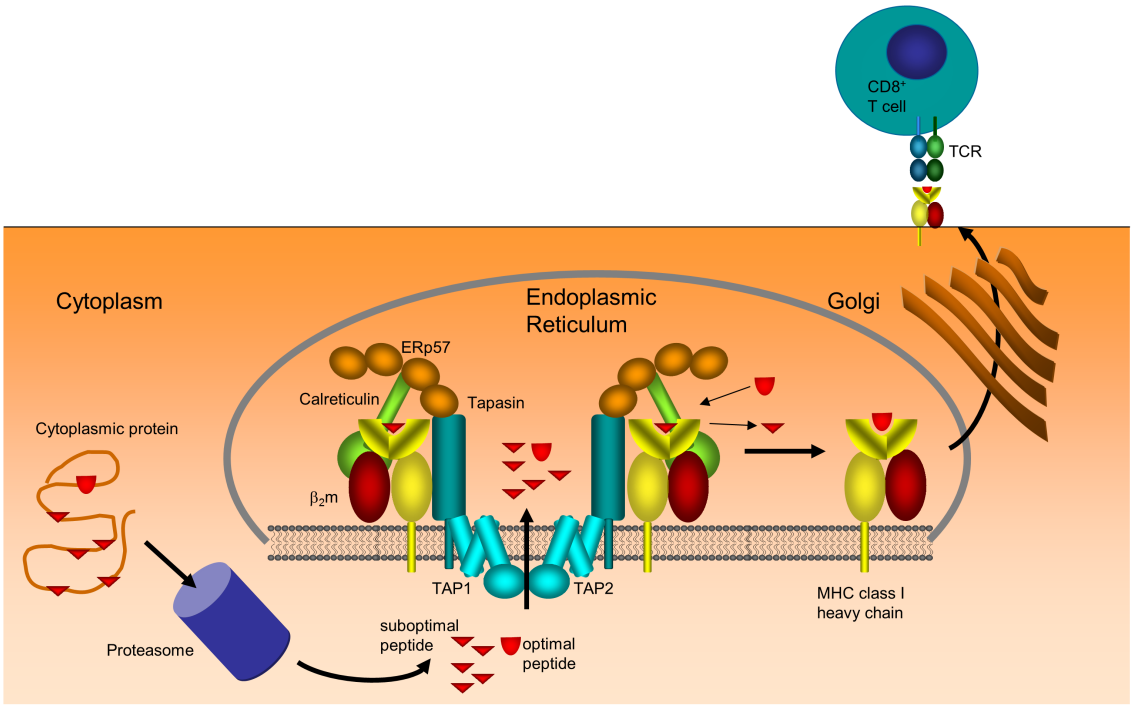


Figure 1.1: The antigen processing pathway[22]. The endogenous antigen processing and presentation pathway: Peptides produced via cytoplasmic degradation of peptide by the proteasome are transported in to the ER via TAP where they are available for binding to MHC class I molecules. The chaperon molecules tapasin, calreticulin, ERp57 and β2m make up the peptide loading complex along with TAP and facilitate the binding process. Figure reproduced from [22].

HIV progression. Such T-cell vaccines have been successful at reducing the viral load of Simian Immunodeficiency Virus (SIV) infected macaques by 1000 fold compared to unvaccinated SIV-infected macaques. However, a successful HIV T-cell vaccine has not yet been achieved.

A patient's immune system could be manipulated to focus on specific disease related epitopes that it otherwise would have been unable to recognise. For example, Toes et al.[25] observed strong and protective tumor-reactive CD8+ T-cell responses when vaccinating a murine model with immunogenic T-cell epitopes from oncogenes required for tumor growth. Alternatively, identification of immunogenic epitopes from self-proteins can be used for de-selection purposes when treating autoimmune diseases[26].

1.3 Modelling Complex Biological Systems

The biological processes occurring in the cells and tissues of the human body are highly complex, with a large number of components and interactions. To be able to understand

such complex systems it is necessary to not only identify the components of the system but also the interactions between them. Systems biology is the study of the interactions within complex biological systems with the aim of understanding the underlying mechanisms of the molecules and pathways. This is important for treating complex diseases such as cancer and Human immunodeficiency virus (HIV).

However, when dealing with complex systems it can be difficult to gain an intuitive understanding of the behaviour of the system and what one would expect to observe if certain aspects were to change. Therefore, the development of computational mathematical models to describe complex systems is becoming more and more common in both biology and medicine, and is a growing area of research[27].

A mathematical model embodies the current hypothesis and knowledge of the system being studied, and by comparing the output with experimentally collected data, any discrepancies or inconsistencies in the dynamic behaviour of the system can be identified and investigated. In this way mathematical models can help support or disprove current beliefs and point to areas where further research is needed. Models which are shown to be consistent with the current knowledge and behaviour of the system can be used as predictions to guide future experiments.

To construct a model of a biological system at the macroscopic level, where we are interested in the concentrations of molecules, a set of equations describing the interactions occurring between the component molecules are required that quantitatively represents each interaction by numerically describing each process using rate constants. Ideally, these rate constants would be determined via experiment, however in many cases the methods or data required are not available, and so the values of the rate constants have to be estimated using fitting techniques or by using the literature to estimate a range in which such a value would fall. No matter how the parameter values are obtained, however, there will always be some magnitude of uncertainty, due to experimental error, lack of data or poor fit of the model to the data. Therefore, there will always be an uncertainty associated with the output of the model, where in this case the output will be the concentrations of each species in the model, either absolute or relative to one another. A sensitivity analysis can reveal which parameters have the biggest impact on the

output of the model and thus determine which parameter values are the source of the most uncertainty. Furthermore, a sensitivity analysis can reveal which are the most important mechanistic steps in a biological system or pathway and so identify possible therapeutic targets.

The emergence of 'high-throughput' approaches has allowed the automation of experiments so that many observations can be carried out in parallel, thus greatly increasing the rate of data collection[28] and providing a wealth of data with which to build computational models of cellular processes. High-throughput experimental techniques are used in several fields, such as genomics, transcriptomics and proteomics and the data collected can then be analysed by bioinformaticians and statistions. For example, in transcriptomics, using high-throughput technology it is possible to measure the expression levels of the mRNA of thousands of genes in parallel, using complementary DNA (cDNA) microarrays[29].

Hondowicz et al. 2012[30] developed a high-throughput protocol to screen CD8+ T-cells, using amplification and *in vitro* transcription of minigene libraries and Enzyme-Linked ImmunoSpot (ELISPOT) assay and identified two novel T-cell epitopes targeted in newly diagnosed type I diabetes sufferers. This demonstrates how high-throughput technologies can be used for T-cell antigen discovery.

Harndahl et al. 2011 [31] used a high-throughput pMHC dissociation assay, screening hundreds of peptides binding to MHC-I molecule HLA-A*02:01. Such data could be used, for example, to help train the machine learning algorithms mentioned in Section 1.3.2 and discussed in more detail in Section 2.2.

Mathematical and computational modelling has been successfully applied to a wide range of biological and medical problems. Extensive research has been carried out in to heart disease by modeling blood flow and electrical activity in the heart with the aim of aiding physicians with treatment options, such as determining the optimal bypass construction[32]. One important area of model development is in describing different aspects of the human immune system. For example, a simple system of two ordinary differential equations (ODEs) can capture the primary and secondary response of the immune system to a target population of pathogen e.g. virus, bacteria or tumour cells [33].

Wodarz & Nowak 2002[34] applied a basic model of viral dynamics to HIV infection, simulating the impact of anti-viral therapies on viral load and the dynamics of antigenic escape using a simplified model of virus-immune interactions.

1.3.1 Modelling the TCR

Considerable effort has gone in to modelling TCR triggering, the earliest of which was McKeithan's 1995[35] paper in which TCR activation was modelled using 'Kinetic proofreading' as had been previously applied to models of DNA replication and protein synthesis[36, 37]. This model is based upon the fact that there is a time lag between the pMHC-TCR binding and T-cell signalling. The model starts off with the TCR in the inactive state, and when the pMHC binds it is required that it stay associated with the TCR long enough for N sequential modifications in the path to signal transduction have been completed. Therefore pMHC with a low dissociation rate are more likely to result in a T-cell signal than those with a high dissociation rate. Those complexes with a high dissociation rate can overcome this by being present in large quantities and so increase the probability of a successful full signal. Dushek et al. 2009[38] modified this simple model to investigate the impact of pMHC-TCR rebinding, whilst van den Berg 2002[39] also modified the model to study T-cell antagonism by peptides that are variants of an immunogenic peptide, which suggest that in addition to dissociation rate, the pMHC-TCR affinity and the pMHC presentation level are also important for T-cell activation. Coombs et al. 2002[40] incorporated this model in to their study of pMHC-induced TCR down-regulation, and the trade-off between kinetic proofreading and serial engagement (where T-cell activation requires a single peptide serial engages multiple multiple TCRs).

Germain and colleagues introduced kinetic proofreading with positive and negative feedback loops, where the negative feedback reduces the number of subsequent activation steps and the positive feedback enhances further signalling[41, 42, 43], leading to bistability in the T-cell response. Predictions made by the model were also verified when Altan-Bonnet and Germain 2004[43] measured ERK-1 (extracellular signal-regulated kinase, the mediator of positive feedback) response at different pMHC densities, and found that - as the model predicted - there is a sharp pMHC-discrimination threshold and so a sharp drop in response time as the number of pMHC decreases.

Wylie et al. 2007[44] adapted this feedback model in an attempt to reconcile the sensitivity of T-cells discrimination between foreign antigen and 'self' with the inhibition of signalling by antagonists (i.e. where certain pMHC complexes reduce T-cell signalling). This study highlighted the importance of accounting for the stochastic nature of biochemical reactions.

Whilst the TCR models which include feedback loops are an improvement on just a simple kinetic proofreading model, there are still major issues with parametrisation, how to model stochasticity and they fail to take in to account the impact of TCR clustering and signal segregation (i.e. spatial effects)[44, 45, 46]. Furthermore, kinetic proofreading models do not consider the mechanisms by which the signal from the TCR is initiated following pMHC binding, but just assume that the signal begins as soon as binding occurs.

Current experimental evidence suggests that TCR triggering occurs following a conformational change in the TCR[47, 48], however the exact mechanism is still not understood. Ma et al. 2008[49] postulate a TCR deformation model in which the mechanical stress at the TCR-pMHC interface results in conformational change of TCR/CD3 complex (where CD3 is a T-cell co-receptor) in an effort to provide a mechanistic explanation for the sensitivity and specificity of TCR triggering, an idea which has gained experimental support[50, 51, 52].

Molecular aggregation models are another class of TCR triggering models, due to experimental evidence suggesting that aggregation of TCR via pMHC oligomers leads to TCR triggering[53, 54]. However, experimental evidence that pMHC oligomerisation actually occurs on the APC is lacking[47].

Segregation models are another class of TCR triggering model that propose TCR triggering is the result of segregation of the TCR/CD3 complex[55]. Burroughs et al. 2006[56] model TCR segregation using a stochastic model of TCR diffusion in the presence of pMHC. Whilst this model yielded some interesting results, it was highly sensitive to certain parameters and overly simplified the kinetic proofreading steps. For a full review of this model, and those mentioned above see [57].

To date there does not exist a model of TCR signalling that can account for all aspects of the process and fits all experimental evidence. It is likely that, since TCR signalling is

a complex and multi-faceted process, some combination of the models discussed above will be required to reconcile all the shortcomings on the models when used on their own.

1.3.2 Predicting peptide-MHC affinity

T-cell epitopes can be identified using high-throughput mass spectrometry methods, however, these methods can only scan a limited number of peptides and MHC-I alleles at a time. Therefore, coupled with the expense of such procedures, it is infeasible at present to perform full scans of all potential T-cell epitopes for complex viruses such as HIV[58].

The peptide-MHC binding step is thought to be the most restrictive in the antigen processing pathway, leading to the development of biochemical competition assays to measure the associated affinity[59], in which the affinity or IC50(nM) value (half the maximal inhibitory concentration), is found by determining the concentration of test peptide required to fill half of the MHC binding sites when competing against a radioactively labelled peptide. By automating such measurements, large numbers of affinities can be measured efficiently, and the Immune Epitope Data Base (IEDB)[60] contains a collection of over $100,000$ measured affinities.

However, the biochemical assays used to measure these affinities require large amounts of resources and equipment which is still quite expensive, and the sheer number of possible peptide sequences for most pathogens, makes scanning of entire genomes infeasible. This has lead to the development of models which predict the affinity of a peptide sequence for a specific MHC allele. Such models use machine learning algorithms, such as Artificial Neural Networks (see Section 2.3), applied to the large sets of experimentally measured peptide-MHC affinity data already collected, in order to make predictions about which peptides will be immunogenic, such as BIMAS[61] and NetMHC[62].

Such methods are used to narrow down the number of peptides screened in high-throughput mass spectrometry experiments. For example, Farrell et al. 2016[63] used the MHC class II binding prediction methods TEPITOPEpan[64] and NetMHCIIPan[65] to select for 376 peptides out of more than a million possible peptides of *Mycobacterium bovis* to be sequenced to screen for promiscuous immunogenic epitopes (i.e. peptides that bind a wide range of MHC alleles and elicit at T-cell response).

However, high affinity does not always correlate with high cell surface abundance,

or immunogenicity, as many other factors also influence the immunogenicity of a peptide. The Immune Epitope Data Base (IEDB)[60] MHC class I processing tool combines peptide-MHC affinity, TAP affinity and proteasomal cleavage probabilities, and provides a measure of how immunogenic a peptide is. The tool combines these three measures into a 'Total Score', which is designed to be proportional to cell surface abundance of the peptide, where the higher the score, the more immunodominant the peptide.

Despite including some of the most important steps in antigen processing, these machine learning algorithms still only provide a static prediction of immunogenicity as they are unable to account for the kinetics of the protein from which the peptide is cleaved, which will influence the availability of peptides for MHC binding.

1.3.3 Models of antigen presentation

Chang et al. 2005[66] constructed a mathematical model of MHC-II antigen presentation within a single macrophage, made up of a set of ODEs describing the levels of MHC-II mRNA, the number of MHC-II both intracellular and on the surface, free self-peptide and self-peptide-MHC-II both surface and intracellular. They also include the impact of IFN-γ by modelling the number of free IFN-γ receptors, the number of free IFN-γ and the number of IFN-receptor-ligand complexes.

Chang et al. use this model to investigate possible mechanisms of inhibition of antigen presentation by *Mycobacterium tuberculosis* (Mtb) using a mathematical model of antigen presentation in macrophages. There are several proposed mechanisms of Mtb antigen presentation inhibition. Moreno et al. 1988[67] proposed inhibition occurs at during antigen processing, Hmama et al. 1998[68] postulated that Mtb down-regulates MHC-II during maturation or peptide loading, whilst Noss et al. 2000[69] suggest that the downregulation occurs at the stage of synthesis of MHC-II mRNA.

By simulating the impact of each of the proposed mechanisms of inhibition, they found that down-regulation of MHC-II is effective at inhibiting antigen presentation in macrophages but there was a time lag of 10 hours between the down-regulation and observing the effects of the inhibition. On the other hand, the impact of the other mechanisms were more immediate, but more easily attenuated. They therefore concluded that the optimal strategy for Mtb antigen inhibition was to target multiple cellular processes.

Dalchau et al. 2011[22] constructed a model of MHC-I antigen presentation, using a set of ODEs to describe the levels of free peptide, MHC and tapasin, the levels of bound pMHC and pMHC-tapasin, and the cell surface levels of pMHC, to investigate process of peptide optimization, where only peptides with a low rate of dissociation will stay bound long enough for the complex to egress to the cell surface. Tapasin has been shown to enhance peptide optimisation[70, 71, 18], and by fitting the model to experimental data, Dalchau et al. showed that tapasin bound to pMHC increases a peptides dissociation rate. They quantified this effect using a *filter relation*, in which the abundance of surface pMHC following the non-tapasin pathway is proportional to $1/u_i$, whereas for the tapasin pathway it is proportional to $1/u_i^2$, where u_i is the pMHC dissociation rate.

1.4 Motivation and Overview

The discovery of T-cell epitopes is of paramount importance in medical science, for the development of T-cell vaccines against certain viruses such as HIV, and for developing therapies to treat cancer and autoimmune diseases. Due to the large number of HLA alleles (over 13,000) and the high degree of polymorphism in MHC renders experimental approaches to identify T-cell epitopes (even high-throughput methods) impractical both in terms of time and expense.

Prediction models built using machine learning techniques can predict peptide specific properties of the antigen processing pathway, such as pMHC binding affinity, proteasomal cleavage probability and TAP affinity. However, these predictions are static as they do not account for protein kinetics, which in turn determines peptide abundance and thus probability of cell surface presentation.

By combining the predictive power of these tools with a mechanistic model of antigen presentation could provide dynamic quantitative predictions of pMHC surface presentation, and thus reduce the number of possible candidate peptides to be screened experimentally.

In this thesis we focus on modelling the intracellular (endogenous) antigen presentation pathway at the molecular level, simulating the dynamics of proteins involved in this subsystem of the immune system, using the MHC-I antigen presentation model presented

in Dalchau et al.[22]. A more rigorous description of the model can be found in Section 2.6.

In Chapter 2 we describe the theory behind mathematical modelling including deterministic and stochastic differential equations, machine learning techniques and the peptide filtering model of Dalchau et al. 2011[22]. In Chapter 3 we extend the peptide filtering model by calibrating it to experimental data of competition between a target and competitor peptide, and use the calibrated model to predict the competition between the target peptide and a range of competitor peptides with different off-rates. In Chapter 4 we simulate the HeLa cell peptidome using experimentally measured HeLa cell protein copy numbers and half-lives. Then, in Chapter 5 we predict the cell surface presentation of HIV epitopes by HLA alleles associated with control and progression of the disease by combining the peptide filtering model with existing models of HIV intracellular kinetics, and using machine learning tools to estimate relative peptide-specific parameters. The general conclusions of this work can be found in Chapter 6.

Chapter 2

Theory

2.1 Introduction: Chemical Reaction Networks

Chemical Reaction Network Theory was developed by Fritz Horn, Martin Feinberg and Roy Jackson[72, 73], and is a framework for modelling how a chemical system evolves in time. The concentrations of the species within a system change as a result of simultaneously occurring chemical reactions, and so in CRNT the aim is to represent the system using a set of differential equations that describe how each species changes in time, and solve these equations to track the concentration of each species over time. The state of the system is then given by the concentrations of each chemical species at any given time. Several assumptions are made in CRNT, such as the concentrations of species cannot be negative, increasing species concentration increases the rate of the reactions it is involved in and the temperature of the system is constant, as is the pressure.

A Chemical Reaction Network (CRN) is a set of chemical reactions $\mathscr{R} := \{R_i | i \in \{1,2,..,n_r\}\}$, between a set of species $\mathscr{S} := \{S_j | j \in \{1,2,..,n_s\}\}$, where any linear combination of the species forms set of complexes $\mathscr{C} := \{C_k | k \in \{1,2,...,n_c\}\}$. We can then define a chemical reaction network as $\mathscr{N} = (\mathscr{S}, \mathscr{R}, \mathscr{C})$. Each reaction can be denoted as:

$$R_i : \sum_{j \in \mathscr{S}} \alpha_{ij} S_j \rightarrow \sum_{j \in \mathscr{S}} \beta_{ij} S_j \tag{2.1}$$

The coefficients α_{ij} and β_{ij} are *stoichiometry coefficients*, and they describe the amount of reactants (left-hand side of Equation 2.1) and products (right-hand side of Equation 2.1), involved in the reaction. The stoichiometry coefficients can be arranged in

to a *stoichiometry matrix*, defined as:

$$[\Gamma]_{ij} = \beta_{ij} - \alpha_{ij} \tag{2.2}$$

For example, consider the following reaction:

$$S_1 + 2S_2 \rightarrow 3S_3 + S_4 \tag{2.3}$$

The stoichiometry matrix associated with this reaction can be calculated as follows:

$$[\Gamma] = \begin{bmatrix} 0 \\ 0 \\ 3 \\ 1 \end{bmatrix} - \begin{bmatrix} 1 \\ 2 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} -1 \\ -2 \\ 3 \\ 1 \end{bmatrix} \tag{2.4}$$

For a CRN with n_s species and n_r reactions, the stoichiometric matrix $[\Gamma]$ has n_s rows and n_r columns. For example, Equation 2.3 describes a single reaction, $n_r = 1$ with four chemical species, $n_s = 4$, and which gives the 1×4 stoichiometric matrix in Equation 2.4. If the reaction were reversible,

$$S_1 + 2S_2 \leftrightarrow 3S_3 + S_4 \tag{2.5}$$

then in this case $n_r = 2$ and stoichiometric matrix is written as:

$$[\Gamma] = \begin{bmatrix} -1 & 1 \\ -2 & 2 \\ 3 & -3 \\ 1 & -1 \end{bmatrix} \tag{2.6}$$

By modelling the CRN mathematically we can determine how the concentrations of each species, $\vec{c}(t) = [c_1(t), c_2(t), ..., c_{n_s}(t)]^T$, changes over time. If we assume that each species is in high enough concentrations and the solution is sufficiently well mixed, that stochastic effects can be ignored, then we can model the dynamics of the system ordinary differen-

tial equations (ODEs). If however, the concentrations of the species are too small that stochastic effects are large, we must then use stochastic modelling.

2.1.1 Deterministic ODE modelling

A deterministic model is a system for which a given set of initial conditions will always produce the same output. This is because deterministic modelling does not include any random terms. When modelling a CRN, the reactions are represented by ODEs, which consist of variables and parameters that describe how the system changes with time, and the reaction is therefore represented as a continuous process. This is because we are assuming a spatially homogeneous distribution of particles, and so a constant reaction rate. In deterministic modelling, the variables are the concentrations of the species involved in the reactions, and the parameters are the rate constants that describe how fast or slow a specific reactions occurs.

To model the dynamics of the system we need to consider net stoichiometric change following each reaction in the network - described by the stoichiometry matrix given in Equation 2.2 - and the rate function of each reaction, $\vec{r}(t,\vec{c}) = [r_1(t,\vec{c}(t))r_2(t,\vec{c}(t)),...,r_{n_r}(t,\vec{c}(t))]^T$. At this point we have not made any assumptions about the rate functions of each reaction, and so they depend on both time and the concentrations of each species involved in the reaction. The dynamics of the system $\mathscr{N}$ can be written deterministically as:

$$\frac{d\vec{c}(t)}{dt} = \Gamma\vec{r}(t,\vec{c}(t)) \tag{2.7}$$

It is common in deterministic modelling of CRNs to assume mass action kinetics, in which the rate of a reaction is proportional to the concentration of the chemical species involved in the reaction. As an example let us consider the following CRN:

Reaction	Rate
$\emptyset \xrightarrow{k_1} S_1$	$r_1 = k_1$
$S_1 \xrightarrow{k_2} S_2$	$r_2 = k_2 c_1$
$S_1 + S_2 \xrightarrow{k_3} S_3$	$r_3 = k_3 c_1 c_2$
$2S_2 \xrightarrow{k_4} S_4$	$r_4 = k_4 c_2^2$
$2S_2 + 2S_4 \xrightarrow{k_5} S_5$	$r_5 = k_5 c_2^2 c_4^2$

Therefore in mass action kinetics the reaction rates for each chemical reaction R_i are defined as:

$$r_i(\vec{c}) = k_i \prod_{j=1}^{n_s} c_j^{\alpha_{ij}} = k_i c_i^{\alpha_i}, \tag{2.8}$$

where we define $c_i^{\alpha_i} = \prod_{j=1}^{n_s} c_j^{\alpha_{ij}}$. Therefore, we define a mass action chemical reaction network as $\mathscr{N} = (\mathscr{S}, \mathscr{R}, \mathscr{C}, \mathscr{K})$, where $\mathscr{K} := \{k_i > 0 | i = 1, ..., n_r\}$, and the time evolution of a system like this can be written as:

$$\frac{d\vec{c}(t)}{dt} = \sum_i k_i c_i^{\alpha_i} \Gamma_i, \tag{2.9}$$

where Γ_i denotes the i^{th} column of the stoichiometry matrix.

2.1.1.1 Solving ODEs

Once we have put together the system of ODEs that describes a CRN using mass action kinetics, in order to visualise what happens to the species within the network and how their concentrations change with time we must solve the system. For example, consider the reaction:

$$\emptyset \xrightarrow{k} Y \tag{2.10}$$

For $k > 0$ this reaction describes the synthesis or supply of species Y with concentration $y(t)$, and is known as the simple exponential growth model. The ODE describing this reaction is written as:

$$\frac{dy(t)}{dt} = ky(t) \tag{2.11}$$

In order to simulate this simple model we need to solve for $y(t)$. For initial conditions

$y(0) = y_0$ the solution is:

$$\int_{y(0)}^{y(t)} \frac{dy(t)}{y(t)} = \int_0^t k dt \tag{2.12}$$

$$ln\frac{y(t)}{y(0)} = kt \tag{2.13}$$

$$y(t) = y_0 e^{kt} \tag{2.14}$$

Using this solution we can now describe the behaviour of the system at any time point. However, it is not usually possible to solve a system of ODEs analytically, and so numerical methods must be used instead. An example of a simple numerical method is the Euler Method. For an initial value problem (IVP) like the simple exponential growth with $y(0) = y_0$ considered above, the Euler method uses the backwards difference formula to produce a sequence of solutions $y_0, y_1, y_2, ...$, such that each past value y_n can be used to approximate y_{n+1}.

To derive the backwards difference formula, consider the change in the quantity y, Δy, between time t and $t + \Delta$:

$$\Delta y = y(t + \Delta t) - y(t) \tag{2.15}$$

We can calculate the rate of change of this function with time i.e. the derivative, by taking the limit of $\Delta y/\Delta t$, as $\Delta t \to 0$:

$$\frac{dy(t)}{dt} = \lim_{\Delta t \to 0} \frac{y(t + \Delta t) - y(t)}{\Delta t} \tag{2.16}$$

Equation 2.16 is just the definition of the derivative. However, if Δt is finite, we can expand $y(t + \Delta t)$ around $y(t)$, using the Taylor series,

$$y(t + \Delta t) = y(t) + \frac{y'(t)}{1!}\Delta t + O(\Delta t) \tag{2.17}$$

where $O(\Delta t)$ refers to the higher order terms of the Taylor series expansion. If we define Δt as h, which will just be the size of the time step, $y(t + \Delta t)$ as y_{n+1}, and $y(t)$ as y_n, we

can write:

$$y_{n+1} = y_n + hy'_n \tag{2.18}$$

This is the backwards difference formula, where the current value y_{n+1}, is calculated using the previous value y_n. The error associated with this approximation, $O(h)$, is of the order h, and so will get smaller as h (or Δt) gets smaller, meaning a better approximation is achieved for smaller time steps. Therefore, for small h, Equation 2.18 will closely approximate the solution of y given in Equation 2.14.

In this thesis, when modelling large systems, such as the model of HIV intracellular kinetics (Chapter 5), we solve our system of ODEs in MATLAB, using their suite of ODE solvers. Our preferred solver is $ode15s$, which is well suited to stiff models, such as where one has a system where some of the components decay much more rapidly than others. The $ode15s$ solver uses backwards differentiation formulas (BDF) optionally, but its default method is numerical differentiation formulas. The Euler method described above is a one-step method, which only calculates the current function value using only one previous value. To achieve greater accuracy, linear multi-step methods can be used where the solution to the ODE is approximated using a linear combination of the function values at several previous time points. This is achieved by approximating $y(t)$ with an interpolating polynomial and differentiating with respect to time. For problems of order k, Newton's Backwards Difference interpolation can be used to write the backwards difference as,

$$\sum_{m=1}^{k} \frac{1}{m} \nabla^m y_{n+1} - hf_{n+1} = 0 \tag{2.19}$$

where $f_{n+1} = y'_{n+1}$, and ∇ is the backwards difference operator, where $\nabla^0 y_n = y_n$ and $\nabla^m y_n = hf(t_n, y_n)$.

The numerical differentiation methods used by $ode15s$ add a term to the BDF to obtain better stability, as follows,

$$\sum_{m=1}^{k} \frac{1}{m} \nabla^m y_{n+1} - hf_{n+1} - \kappa\gamma_k \nabla^{k+1} y_{n+1} = 0 \tag{2.20}$$

where κ is a scalar parameter and $\gamma_k = \sum_{j=1}^{k} \frac{1}{j}$.

To minimize the time it takes to solve the system, there is an option to provide a

Jacobian, $\mathbf{J} = \frac{d\mathbf{y}}{d\mathbf{x}}$, where $\mathbf{x}$ is the vector of parameters of the system. Otherwise, the solver would approximate the Jacobian, and for large systems such as the HIV model, where many proteins and individual peptides are being tracked, this can make the solver very inefficient.

2.1.2 Stochastic modelling

In deterministic modelling, the ODE representations of the reactions, describes the population of the set of species S (as defined in Section 2.1.1), evolving with time as a whole, and does not take in to account that fact that such a system is actually made up of individual particles. Consider for example the degradation of species Y:

$$Y \xrightarrow{k_{-1}} \emptyset, \tag{2.21}$$

where $\emptyset$ means the species is either leaving the network, or degrading, and so it is no longer part of the network. If we wrote this reaction as a deterministic ODE, we would be making the assumption that degradation of each molecule of Y is coordinated with one another, so that a certain number of them will always degrade within a specified time interval. In reality, however, each molecule of Y will have a half-life that is independently and identically distributed (i.i.d) from within the same distribution, but their values will be mutually exclusive of one another.

Therefore, deterministic ODEs in fact describe the *average* population of Y, and how it evolves with time, and the half-lives of the molecules will be exponentially distributed random variables. In the situation where the number of the reaction molecules is very small, you would expect to see a very large variation in their half-lives from the mean, due to the relatively low number of samples from the distribution. In this case, the deterministic approach would not capture the behaviour of the system very well. Stochastic modelling therefore deals with the probabilities of a reaction occurring within a certain time-frame, and the exact number of molecules present in the system at that time. Such a system is in turn modelled using differential equations where the probabilities are the variables.

If we consider again CRN as defined in Section 2.1 with n_s species of molecules that

can undergo n_r different types of reactions involving a subset of the constituent molecular species, with the number of each species given by the vector $\vec{c}(t)$. Each of these n_r possible reactions will change the number $c_j(t)$ of the species involved by an integer amount. Therefore, as before for each reaction $R_i \in \mathscr{R}$ we can determine the change in the state vector using the stoichiometric matrix Γ, as defined in Equation 2.2.

In stochastic theory, it is assumed that transitions in the state space occur randomly in time as the result of accidental collisions between the set of species $\mathscr{S}$, i.e. the events are continuous and independent of one another. This is known as a Poisson process, and each reaction is assumed to occur according to the exponential probability distribution. The rate at which a reaction occurs is a function of only the current state of the system, in a Markov-like way. Therefore, the likelihood of reaction $i \in \{1,2,...,n_r\}$ occurring in time interval Δt is defined by a *propensity function*, $a_i(\vec{c}(t))$ and so the probability that reaction i will occur in the small time interval $t+\Delta t$ is $a_i(\vec{c}(t))\Delta t$.

We therefore construct a system of ODEs to describe the probability of the system being in a certain state (i.e. specific integer number of molecules of all of the n_s species, described by $\vec{c}(t)$) at time t. The number of possible differential equations is the number of possible states, n, that the system can be in, and this collection of differential equations is known as the Chemical Master Equation (CME)[74].

For each possible state we construct the ODE on the probability of the system being in that state n with species numbers described by $\vec{c}(t+\Delta t)$ at time $t+\Delta t$. We do this using the law of total probabilities. There are only two possible ways that the system can come to be in state n at time $t+\Delta t$, either it arrives in state n at time $t+\Delta t$, or it was in state n at time t and has remained in that state at time $t+\Delta t$, i.e. $\vec{c}(t+\Delta t) = \vec{c}(t)$.

If the system arrives in state n at time $t+\Delta t$, then this means that the previous state of the system was exactly $\vec{c}(t+\Delta t) - \Gamma_i$ for all reactions i. Therefore, the probability that the state arrives in state n at time $t+\Delta t$ is given by,

$$P_{arrive} = \Delta t \sum_{i=1}^{n_r} a_i(\vec{c}(t+\Delta t) - \Gamma_i) P(\vec{c}(t+\Delta t) - \Gamma_i, t) \tag{2.22}$$

where $P(\vec{c}(t+\Delta t) - \Gamma_i, t)$ is the probability that the system was in a state where the number of molecules was exactly $\vec{c}(t+\Delta t) - \Gamma_i$ at time t.

If the system was in state n at time t and has remained in state n at time $t+\Delta t$, then the probability that it remained is simply one minus the probability for it to leave,

$$P_{stay} = \left[1 - \Delta t \sum_{i=1}^{n_r} a_i(\vec{c}(t))\right] P(\vec{c}(t), t) \tag{2.23}$$

where $P(\vec{c}(t), t)$ is the probability that the system was in state n with number of molecules $\vec{c}(t)$ at time t.

Therefore, the probability that the system is in state n with number of molecules $\vec{c}(t+\Delta t)$ at time $t+\Delta t$ is,

$$\begin{aligned} P(\vec{c}(t+\Delta t), t+\Delta t) &= P_{arrive} + P_{stay} \\ &= \Delta t \sum_{i=1}^{n_r} a_i(\vec{c}(t+\Delta t) - \Gamma_i) P(\vec{c}(t+\Delta t) - \Gamma_i, t) + \left[1 - \Delta t \sum_{i=1}^{n_r} a_i(\vec{c}(t)\right] P(\vec{c}(t), t) \end{aligned} \tag{2.24}$$

Rearranging Equation 2.24 we get:

$$\frac{P(\vec{c}(t+\Delta t), t+\Delta t) - P(\vec{c}(t), t)}{\Delta t} = \sum_{i=1}^{n_r} a_i(\vec{c}(t+\Delta t) - \Gamma_i) P(\vec{c}(t+\Delta t) - \Gamma_i, t) - a_i(\vec{c}(t)) P(\vec{c}(t), t) \tag{2.25}$$

Taking the limit where $\Delta t \to 0$, we get the Chemical Master Equation:

$$\frac{dP(\vec{c}(t), t)}{dt} = \sum_{i=1}^{n_r} a_i(\vec{c}(t) - \Gamma_i) P(\vec{c}(t) - \Gamma_i, t) - a_i(\vec{c}(t)) P(n, t) \tag{2.26}$$

Therefore, the current state of the system, only depends upon the state directly preceding it and not any of the other historic states, meaning the CME is a Markov process.

2.1.2.1 Stochastic simulation algorithm

For very large systems, the CME is very difficult to solve, and so instead the Stochastic Simulation Algorithm (SSA), also called Gillespie's Algorithm, is used where realisations of the state of the system are computed in accordance with the probability distribution described by the CME.

We want to compute the time $t+\theta$, when the next reaction occurs and which reaction it will be. The first step in the SSA is to determine the probability that the ith reaction

occurs in time interval $[t+\tau, t+\tau+\delta\tau)$, denoted $P(\tau, i|\vec{c}(t), t)\delta\tau$. To do this first lets consider the probability of the system being in state n with $\vec{c}(t)$ molecules and no reaction occurring in interval $[t, t+\tau)$ which we will name $P_0(\tau|\vec{c}(t), t)$. If we want to compute the probability that no reaction occurs over the interval $[t, t+\tau+\delta\tau)$, we can therefore write this as the probability of no reaction occurring over interval $[t, t+\tau)$, $P_0(\tau|\vec{c}(t), t)$, multiplied by the probability of no reaction occurring over interval $[t+\tau, t+\tau+\delta\tau)$ (i.e P_{stay} given in Equation 2.23). Therefore,

$$P_0(\tau+\delta\tau|\vec{c}(t), t) = P_0(\tau|\vec{c}(t), t)\Big(1 - \sum_{i=1}^{n_r} a_i(\vec{c}(t))\Big)\delta\tau \tag{2.27}$$

As before, if we rearrange we get:

$$\frac{P_0(\tau+\delta\tau|\vec{c}(t), t) - P_0(\tau|\vec{c}(t), t)}{\delta\tau} = -\sum_{i=1}^{n_r} a_i(\vec{c}(t)) P_0(\tau|\vec{c}(t), t) \tag{2.28}$$

Taking the limit where $\delta t \to 0$ and solving for $P_0(\tau|\vec{c}(t), t)$:

$$P_0(\tau|\vec{c}(t), t) = exp(-\sum_{i=1}^{n_r} a_i(\vec{c}(t))\tau) \tag{2.29}$$

This shows that the time to the next reaction when in state n is exponentially distributed with mean value $1/a_0$, where $a_0 = \sum_{i=1}^{n_r} a_i(\vec{c}(t))$. Next we will consider the probability that the next reaction to take place over interval $[t+\tau, t+\tau+\delta\tau)$ is of type i, which is simply dependent upon the propensity function $a_i(\vec{c}(t))\delta\tau$. Therefore, the probability that the ith reaction occurs in time interval $[t+\tau, t+\tau+\delta\tau)$, $P(\tau, i|\vec{c}(t), t)\delta\tau$, is given by the product of the probability that no reaction occurred in interval $[t, t+\tau)$, $P_0(\tau|\vec{c}(t), t)$, and the probability that the ith reaction occurred in interval $[t+\tau, t+\tau+\delta\tau)$:

$$P(\tau, i|\vec{c}(t), t) = P_0(\tau|\vec{c}(t), t) a_i(\vec{c}(t)) = a_i(\vec{c}(t)) exp(-\sum_{i=1}^{n_r} a_i(\vec{c}(t))\tau) \tag{2.30}$$

To simulate the SSA we require the time to the next reaction and the index of the next reaction. We re-write equation 2.30 as:

$$P(\tau, i|\vec{c}(t), t) = \frac{a_i(\vec{c}(t))}{\sum_{i=1}^{n_r} a_i(\vec{c}(t))} \cdot \sum_{i=1}^{n_r} a_i(\vec{c}(t)) exp(-\sum_{i=1}^{n_r} a_i(\vec{c}(t))\tau) \tag{2.31}$$

We can then define the *next reaction index* as $\frac{a_i(\vec{c}(t))}{\sum_{i=1}^{n_r} a_i(\vec{c}(t))}$, where the probability of the next reaction being the ith reaction is therefore the relative contribution of $a_i(\vec{c}(t))$ to the total reaction rate of the system i.e. for each reaction i, the next reaction index is a random variable s_1 which lies in the uniform interval $[0,1]$. We can also say that $\sum_{i=1}^{n_r} a_i(\vec{c}(t))exp(-\sum_{i=1}^{n_r} a_i(\vec{c}(t))\tau)$ is the density function of exponentially distributed random variable τ with parameter $\sum_{i=1}^{n_r} a_i(\vec{c}(t))$. It can be shown that the time to the next reaction can be written as,$\tau = ln(1/s_2)/a_0$, where s_2 is a random number in the uniform interval $[0,1]$. Therefore, $P(\tau, j|\vec{c}(t),t)$ is the joint probability distribution for the two random variables i and τ.

The Gillespie algorithm is based upon these two observations. In order to compute realisations of the state of the system, the reaction index s_1 and reaction time s_2 can be independently sampled from the uniform distribution $[0,1]$. The steps of the algorithm are given below in Algorithm 1.

Algorithm 1 Gillespie's Stochastic Simulation Algorithm

1. Initial the state of the system $n_0 = n$ at $t_0 = t$, and the total reaction propensities in that states, $a_0 = \sum_{i=1}^{n_r} a_i(\vec{c}(t))$.
2. Draw two independent random numbers s_1 and s_2 from the uniform distribution $[0,1]$.
3. Determine the index of the next reaction by taking the smallest value of i for which $\sum_{k=1}^{i} a_k(\vec{c}(t)) = s_1 a_0$.
4. Determine the time to the next reaction by calculating $\theta = ln(1/s_2)/a_0$.
5. Update the state of the system: $\vec{c}(t+\theta) = \vec{c}(t) + \Gamma_i$.
6. Update time $t \to t + \theta$.
7. Re-evaluate the total reaction propensities for the new state, $a_0 = \sum_{i=1}^{n_r} a_i(\vec{c}(t+\theta))$.
8. Go to step 2.

The algorithm terminates when a threshold of time has been simulated.

2.2 Machine learning

Machine learning is the name of an area in computer science where algorithms are defined by 'training' them on existing sets of data. The simplest example of this is a classifier, where when an input vector is provided it will produce a single discrete output. For

example, a simple model for predicting whether a peptide binds to an MHC allele, would take the sequence of amino acids of the peptide as input and output either 'binding' or 'non-binding'.

In order to create such a predictive model, the parameters of the classifier must first be determined by training the model on a set of data. The model will 'learn' which sets of inputs are most likely to produce a certain output, and produce a generalised predictive model. The more data available with which to train the model, the better the model will be at correctly predicting the class of input data not included in the training set. This type of machine learning is called Supervised Learning, where the algorithm learns a set of rules to determine the output from the input, and it can also be used to produce continuous outputs, such as predicting the IC50 of a peptide to an MHC allele.

Other types of machine learning include Unsupervised Learning and Reinforcement Learning. In Unsupervised Learning, the desired output is not provided, but the algorithm must find some structure within the input data, and identify similarities or differences between them. In Reinforcement Learning, the algorithm must optimise itself through trial and error to produce the best decision pathway to achieve the best result. This type of machine learning is often used in robotics and navigation.

Machine learning has been used to predict peptide binding to MHC alleles, and there are several algorithms available online, where one can input a protein or peptide sequence and the model will output a result pertaining to how well the peptide binds, such as the IC50, the half-life, or a numerical score which predicts binding to occur over a certain threshold value.

For example the BIMAS (BioInformatics and Molecular Analytics Section) predictor[61], which can be found at `http://www-bimas.cit.nih.gov/molbio/hla_bind/`, predicts the half-time of dissociation of a peptide from an HLA allele, whereas the IEDB (Immune Epitope Data Base) predictor[75], which can be found at `http://tools.iedb.org/mhci/`, produces a predicted IC50 (nM) value for the peptide to the HLA allele.

To predict the HIV peptidome, we used the IEDB prediction tool, as it is able to produce predictions for many more HLA alleles than the BIMAS predictor, and so enables

us to compare between several known controlling and non-controlling alleles. The IEDB predictor allows the user to choose between several different prediction methods. We chose the IEDB recommended Consensus method[76], which combines Artificial Neural Networks (ANN)[77, 78], Stabilized Matrix Methods (SMM)[79] and Scoring Matrices derived from Combinatorial Peptide Libraries (Comblib)[80].

These three methods are combined as it is thought that combining the output of several predictive methods may lead to a better predictive method overall[81]. The predictions from each of the three predictors are converted in to percentile rank scores, and the overall score for a peptide is taken as the median of the three percentile rank scores[60].

2.3 Artificial Neural Networks

Artificial neural networks (ANN) are computational models that are based upon biological neurons in the human brain[82]. There are three layers in ANNs: input, hidden and output (Figure 2.1 a) and each layer is made up of neurons, or nodes. The weighted sum of the inputs to a node is calculated, and the output of a certain node o_i is determined by the applying an activation function to the weighted sum (Figure 2.1 b). For a given set of input values $x_1, x_2, ..., x_n$, there will be a corresponding set of output values, such as y_1, y_2 for a network with only two output nodes, for which there is either a known or desired value.

When training a neural network, the connection weights $\omega_{i,j}$, that connects node i to node j, are set at random, and are then modified during the training, using a method known as 'back-error propagation'. Each time the system is run, there will be an error between the output of the system t_1, t_2 for instance, and the required output values y_1, y_2. Back-error propagation aims to minimise this error, by using the gradient descent method to calculate the gradient of the error function with respect to all the weights in the neural network. A node with k inputs will have a corresponding $k+1$-dimensional error surface, and the gradient function will find the minimum of that surface by taking the derivative and following the direction of maximum negative gradient, and updating the weighting parameters accordingly.

The back-error propagation method requires the activation function to be differen-

tiable. The activation function represents the 'firing' of the neuron. The simplest activation functions are binary, and so the neuron is either firing or not firing. However, for multi-layer networks, non-linear logistic classifiers work better, which output a probability between 0 and 1 that the neuron is firing. The most commonly used activation function of this type is the sigmoidal function, $\phi(z) = 1/(1+e^{-z})$, as shown in Figure 2.1 c.

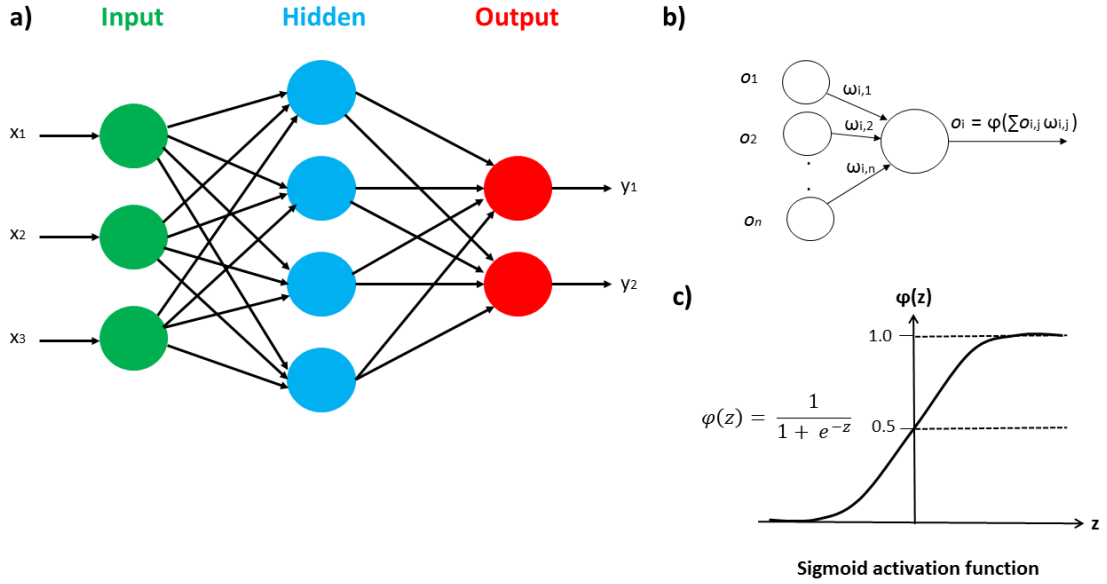


Figure 2.1: Artificial neural networks (ANN) are based upon biological neurons. a) The three layers in ANNs: input, hidden and output. b) The output of a node is determined by applying an activation function (c) to the weighted sum of the inputs to a node.

2.4 Stabilized Matrix Method

The Stabilizied Matrix Method (SMM), proposed by Peters et al. 2003[79], is a modification to the assumption that each peptide in a sequence contributes independently contributions to binding (AIB) to MHC-I, by including pair-wise interactions between the amino acids within the peptide and also considering the errors inherent in any experimental data, to avoid over-fitting.

In the SMM method, they first construct a scoring matrix to quantify the contribution of each amino acid residue within the peptide to the binding affinity of the peptide to the MHC allele. This matrix assumes independent contributions to binding, and so the contribution of amino acid A and position i within the peptide, is given by $s_{A,i}$, and thus

the total score S_k for a peptide k is given by:

$$S_k = s_0 + \sum_i s_{A,i} \tag{2.32}$$

where s_0 is a constant offset.

During training, the aim is to minimize the difference between S_k and the measured IC50 value. However, in order to avoid over-fitting, a second term is added to the minimization function, which introduces a trade-off between fitting to the data, whilst also minimizing the parameters $s_{A,i}$:

$$\Psi(\{s_{A,i}, \lambda\}) = \sum_k ||S_k - data_k|| + \lambda \sum_{A,i} s_{A,i}^2 \tag{2.33}$$

The impact of the second term on the right hand side is to avoid over-fitting to the data, by ensuring that the fitting process is biased towards those values of $s_{A,i}$ which are important to the minimisation of the first term on the right hand side of the equation i.e., those values of $s_{A,i}$ which do not significantly improve the fit to the data are forced towards zero. To incorporate the interactions between the amino acid A at position i, and amino acid A' at position i' within a peptide, pair-coefficients $s'_{A,i,A',i'}$ are introduced in to the total score calculation. Therefore, for peptide k,

$$S'_k = S_k + \sum_i \sum_{i'} s'_{A,i,A',i'} \tag{2.34}$$

and the minimization function becomes:

$$\Psi'(\{s'_{A,i,A',i'}, \lambda'\}) = \sum_k ||S'_k - data_k|| + \lambda' \sum_{A,i,A',i'} s'^2_{A,i,A',i'} \tag{2.35}$$

The pair-coefficient $s'_{A,i,A',i'}$ quantifies the difference in the binding affinity of a peptide k when amino acid A is at position i AND amino acid A' is at position i' compared with the independent average contribution to the binding affinity when amino acid A at position i i.e. $s_{A,i}$ and amino acid A' at position i' i.e. $s_{A',i'}$. Including the pair interactions improves the predictions compared to the AIB method, and the SMM predictions outper-

forms Gulukota et al.'s ANN method[83], Segal et al.'s classification tree method[84], and Doytchinova et al.'s additive method[85].

2.5 Combinatorial Peptide Library

A combinatorial peptide library is a mixture of a very large number of different peptides. Sidney et. al[80] present a positional scanning combinatorial library, in which the affinity of a large pool of peptides, all of which share a single residue at a specific position, is calculated to determine the average influence the shared residue has on binding affinity overall. For a 9-mer where each position can be one of the 20 possible amino acids, there will be 180 (9x20) different pools of peptides. The affinity of each peptide to an MHC alleles is determined by experimentally measuring the IC50 (nM), which in this case is approximately the dissociation constant K_D.

For each position in a peptide, the contribution of each of the 20 residues is determined by calculating the average relative binding (ARB) affinity, under the assumption that each position contributes independently. Each position is associated with a specificity factor SF, which is the amount by which the ARB of all 20 amino acids at that position differs from the average affinity of the entire library. Primary anchors are those positions with a specificity factor SF > 2.4, and secondary anchor positions were identified by determining the standard deviation of the ARB values for each of the 20 amino acids at each position. A peptide is then given a score which is the product of the value for the residue at that position for each peptide of the 9 amino acids in the peptide.

2.6 A Peptide Filtering Model

Much of the work in this thesis is based upon the dynamical systems model of peptide filtering conceived by Dalchau et al. 2011[22], therefore a detailed description of the model is required. Figure 2.2 shows a diagrammatic representation of the peptide filtering model. The change in concentration of each molecular species in the system is assumed to conform to the laws of mass-action kinetics, and so the system can be described by a set of coupled ordinary differential equations (ODEs), as follows:

$$\frac{d[P_i]}{dt} = g_i - d_P[P_i] - b_P[P_i][M] - c[P_i][TM] + u_i[MP_i] + u_i q[TMP_i] \tag{2.36}$$

$$\frac{d[M]}{dt} = g_M - d_M[M] - b_T[M][T] + u_T[TM] - b_P[P_i][M] + u_i[MP_i] \tag{2.37}$$

$$\frac{d[T]}{dt} = g_T - d_T[T] - b_T[M][T] + u_T[TM] + u_T v[TMP_i] \tag{2.38}$$

$$\frac{d[MP_i]}{dt} = b_P[P_i][M] - u_i[MP_i] + u_T v[TMP_i] - e[MP_i] \tag{2.39}$$

$$\frac{d[TMP_i]}{dt} = c[P_i][TM] - u_i q[TMP_i] - u_T v[TMP_i] \tag{2.40}$$

$$\frac{d[MeP_i]}{dt} = e[MP_i] - u_i[MeP_i] \tag{2.41}$$

$$\frac{d[Me]}{dt} = u_i[MeP_i] - d_{Me} \tag{2.42}$$

This model begins with peptide supply, g_i from the cytoplasm in to the ER, where the peptide can either degrade with rate d_P, or bind to an MHC allele with rate b_P or MHC-tapasin complexes with rate c (where $c > b_P$, Equation 2.36). MHC is supplied to the ER at rate g_M, and degrades with rate d_M. Similarly, tapasin is supplied at rate g_T and degrades at rate d_T. MHC and tapasin bind with rate b_T and unbind with rate u_T. When a peptide binds to a tapasin-MHC complex, the presence of tapasin increases the peptide unbinding rate by factor q so only high affinity peptides will stay bound long enough for tapasin to unbind allowing the peptide-MHC complex to egress to the cell surface. Furthermore, the presence of the peptide increases the unbinding rate of tapasin from the tapasin-MHC-peptide complex by factor v. The peptide-MHC complex egresses to the cell surface at rate e. In the peptide filtering model, it is assumed the supply rate of each peptide g_i is a constant and does not depend upon the cytoplasmic peptide concentration, or the peptide affinity with TAP. Furthermore, the peptide-MHC binding rate b_P is assumed to be the same for each peptide in the system, and so differences in affinity for MHC are determined by the peptide-MHC unbinding rate, u_i. To reach the cell surface a peptide-MHC complex must leave the ER and pass through the Golgi (see Figure 1.1),

however in the model, the ER and the Golgi are not considered to be two separate compartments, and so the rate e describes the egress out of the ER *and* through the Golgi. Once at the cell surface a peptide-MHC complex can irreversibly dissociate with rate u_i, and the empty MHC will then decay with rate d_{Me}. By considering the equilibrium solution to the system of ODEs (Equations 2.36 - 2.42), Dalchau et al.[22] arrived at the following approximations for the cell surface abundance of a peptide in the absence and presence of tapasin respectively:

$$[MeP_i]^* = N_i \frac{e}{u_i + e} \overset{e<<u_i}{\approx} N_i \frac{e}{u_i} \tag{2.43}$$

$$[MeP_i]^* = N_i \frac{x}{u_i + x} \frac{e}{e + u_i} \overset{e,x<<u_i}{\approx} N_i \frac{ex}{u_i^2} \tag{2.44}$$

In the absence of tapasin (Equation 2.43), the concentration of peptide P_i on the surface in equilibrium, $[MeP_i]^*$, with off-rate u_i, can be approximated by the $N_i e/u_i$ in the limit $e << u_i$, i.e. under the assumption of maximal optimization, where only the most stable pMHC complexes egress to the cell surface. In the presence of tapasin (Equation 2.44), $[MeP_i]^*$ can be approximated by $N_i ex/u_i^2$, where $x = u_T v/q$, this time in the limit $e, x << u_i$, which again assumes only the most stable complexes will egress. Therefore, in the absence of tapasin, the number of peptide-MHC cell surface complexes at equilibrium with unbinding rate u_i is proportional to $1/u_i$, whereas in the presence of tapasin, this number is proportional to $1/u_i^2$. As mentioned in the introduction, the chaperone molecule TAPBPR was only recently discovered to have a similar optimising effect as tapasin by filtering out low affinity peptides The filtering model presented here was put together before the TAPBPR filtering mechanism was fully known, therefore TAPBPR was not included in this model.

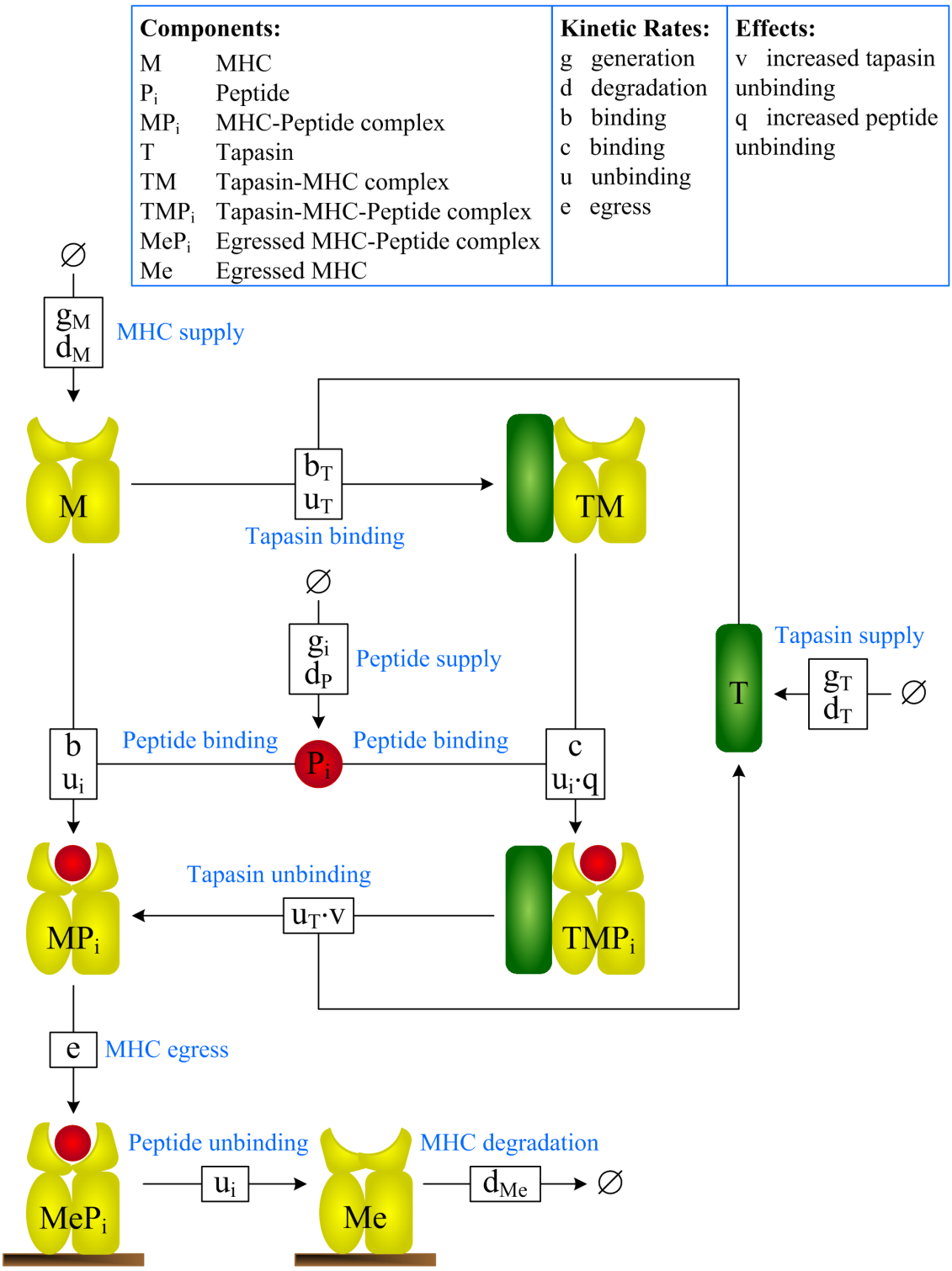


Figure 2.2: A peptide filtering model[22]. Peptides are supplied to the ER via TAP where the rate for each peptide is denoted g_i. Once in the ER peptides can either bind (parameter b_P) and then unbind (parameter u_i) from an MHC allele (M), or the peptide-MHC complex (MP_i) can egress to the cell surface (rate parameter e). The other pathway a peptide can take is to bind to an MHC-tapasin complex (TM) denoted by rate constant c and unbind denoted by parameter $u_i q$. Or, the tapasin can unbind ($u_T v$) before the peptide and then the complex can egress to the cell surface. The MHC allele is produced with rate coefficient g_M and degrades at d_M, and tapasin is produced and degrades with rate parameters g_T and d_T respectively. Tapasin and MHC can bind and unbind, denoted by the rate coefficients b_T and u_T respectively. Figure reproduced from [22].

Chapter 3

A Model to Predict Peptide Competition for MHC class I Binding and Presentation

3.1 Introduction

The first vaccine was developed over 300 years ago when an animal poxvirus was used to vaccinate against smallpox, leading to the eventual worldwide eradication of smallpox. Vaccines commonly use inactivated or attenuated forms of the virus, or in the case of protein vaccines, a highly immunogenic component of the virus, such as the capsid proteins. These types of vaccines induce a humoral response, where antibodies produced by B-cells destroy the extracellular pathogen and so offer antibody-mediated immunity. They do not, however, offer cell-mediated immunity provided by T-cells and B-cells. Attenuated virus and protein vaccines can contain thousands of proteins, however, the immune response is usually only dependent on a small number of the proteins. Such a large number of proteins can cause allergenic and reactogenic responses, are at risk of contamination from extraneous sources, and are inflexible to escape variants[86]. Peptide vaccines, on the other hand, are easily produced, are able to induce T-cell responses and can be easily adapted to escape variants.

As previously mentioned, T-cell vaccines have been successfully used to immunise mice against HIV[23], to vaccinate a murine model against tumour growth[25], and

Rosario et al. used a synthetic HIV peptide vaccine to boost HIV-specific T cell responses in a macaque model[87]. However, the selection of effective peptide sequences to include in a peptide vaccine is complicated by the sheer sequence diversity of even short peptides of between 8 and 11 amino acids. Furthermore, it is difficult to know whether a given sequence will remain abundant at the cell surface for long enough to prime circulating T-cells.

In the search for immunogenic epitopes, the development of accurate and reliable predictive models is of paramount importance in immunoinformatics. However, the machine learning algorithms discussed in Chapter 2, which provide predictions of MHC binding affinity, may not be sufficient on their own, as a high predicted affinity does not always correlate with peptide cell surface abundance or CTL response. For example, Feltkamp et al.[88] found that whilst efficient MHC binding is required for CTL response, it does not always guarantee that a peptide with a high affinity will be immunogenic. Similarly, Ochoa-Garray et al.[89] found a weak correlation between peptide-MHC affinity and CTL immunogenicity. Furthermore, peptide-MHC affinity was also found to have a weak correlation with cell surface abundance[90], a precursor to immunogenicity.

The likelihood that a peptide will be presented on the cell surface also depends on the concentration of peptide in the ER, which in turn depends upon their affinity for TAP, and the abundance and degradation of their source proteins. For example Bassani-Sternberg et al. 2015[91] found that peptide-MHC abundance correlates with both protein abundance and degradation rate. However, Milner et al. 2006[92] reported only a weak correlation between the abundance of peptide-MHC and the abundance of their proteins of origin.

Furthermore, the cytokine interferon gamma (IFN-γ), which is primarily produced as part of the immune systems response to pathogenic infection or cancer, increases the expression of MHC-I proteins, the immunoproteasome, TAP and tapasin, meaning it can profoundly influence the immunopeptidome. Aberrant expression of IFN-γ is strongly linked with the development of systemic autoimmune diseases, such as lupus, and is also less strongly linked to rheumatoid arthritis and multiple sclerosis[93].

Therefore, as no strong correlation has been observed between peptide-MHC abundance and any one of the steps in the antigen processing pathway on its own, it is likely

that a combination of these factors is required to obtain accurate predictions of T-cell epitopes. However, the machine learning approaches developed thus far produce static predictions that do not incorporate peptide abundance, and therefore will under-estimate the potential relevance of peptides originating from highly abundant proteins. However, the relative importance of peptide abundance and peptide-MHC stability in determining cell surface abundance is yet to be determined. While the peptide filtering model of MHC class I presentation[22] includes explicit terms for the supply of peptide into the ER (which should scale with the cytosolic abundance of the parent protein), and peptide-MHC stability, the effects of differential peptide abundance have not yet been tested experimentally.

In this chapter, we have directly addressed the question of the relative importance of peptide abundance and peptide-MHC stability. We have augmented the peptide filtering model to interpret experimental measurements that directly measure the intracellular peptide abundance of two different peptides and simultaneous measurements of those peptides bound to MHC class I molecules at the cell surface of individual cells.

The peptide filtering model described in Section 2.6 can be used to simulate the cell surface presentation of a range of peptides with different MHC unbinding rates and ER supply rates[22]. We decided to see if the model could be used to predict the cell surface presentation of two competing peptides with different MHC unbinding rates at a range of different abundances, and in the presence and absence of IFN-γ.

3.2 Experimentally Quantifying Peptide Competition

Our experimental collaborators in the University of Southampton Cancer Science Unit developed an assay to measure the abundance of two competing peptides in the cytoplasm and the cell surface simultaneously, using an assay adapted from Lev et al. 2010[94]. A target peptide SSLENFRAYV (SSL) was endogenously fused with a ubiquitin tagged fluorescent protein (mCherry) and a competitor peptide ASNENMETM (ASN) was similarly tagged with the fluorescent protein (Venus)[95](see Figure 3.1 A). The target peptide SSL had a much slower off-rate ($2.8 \times 10^{-5} s^{-1}$) to the mouse MHC allele H2Db than the competitor peptide ASN ($5.2 \times 10^{-5} s^{-1}$). In the cytoplasm, the tagged peptides were

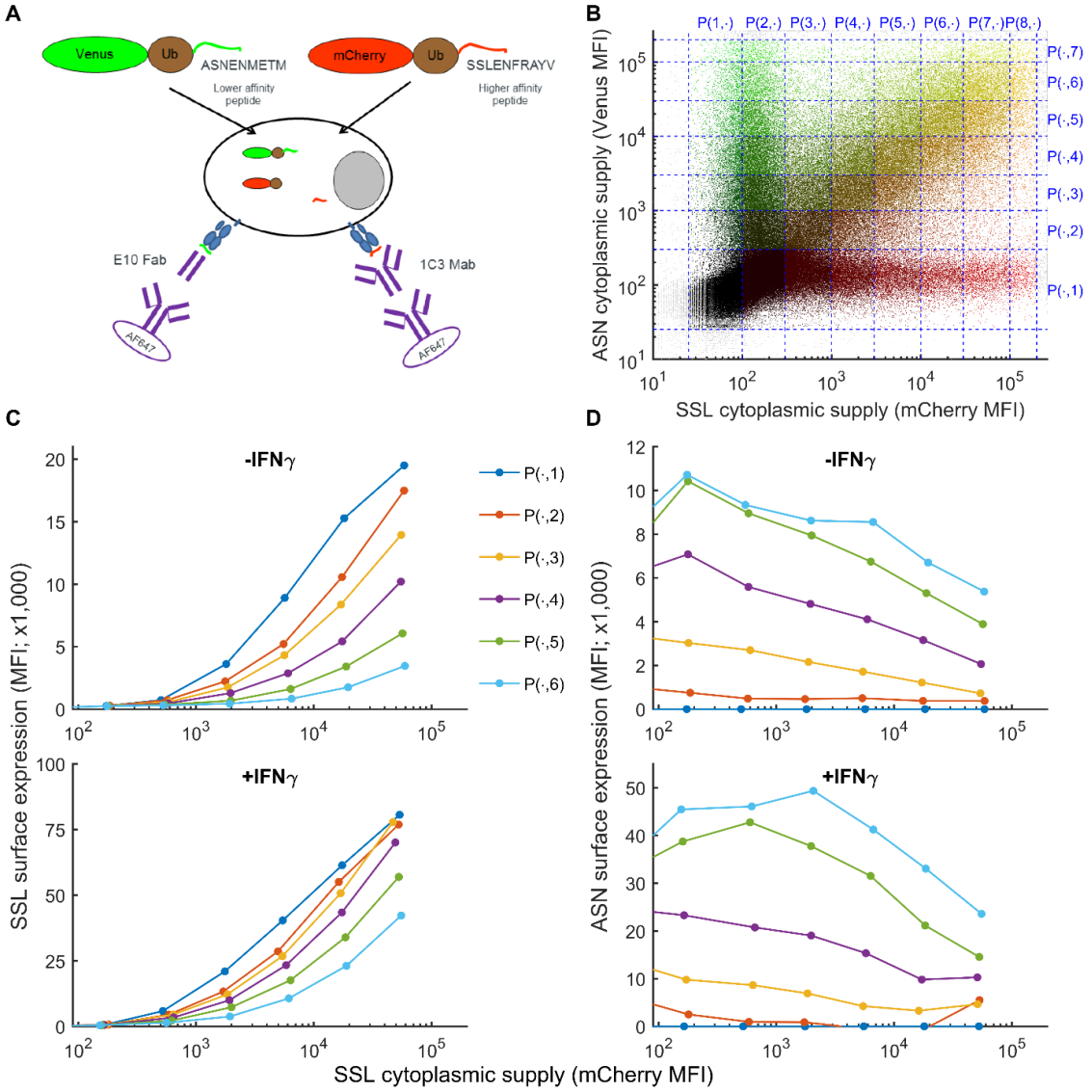


Figure 3.1: Simultaneous measurement of intracellular peptide abundance and cell surface pMHC. (A) Experimental setup used by our collaborators. Fibroblasts were co-transfected with constructs expressing fusion proteins made of a fluorescent protein, ubiquitin and a peptide. Cytoplasmic ubiquitin hydrolases cleave the fusion proteins, releasing an equimolar ratio of peptide and fluorescent protein. ASNENMETM-H2Db surface complexes were be detected using E10 Fab and SSLENFRAYV-H2Db using 1C3 chimeric Mab. (B) In a single transfection assay cells were expressing low to high levels of both fusion proteins and were separated in different gates for the purpose of the analysis. (C) Cell surface expression SSLENFRAYV-H2Db in the presence of increasing competitor concentration. The dark blue curve corresponds to gates P(1, 1) to P(8, 1) with no competitor and shows the maximum surface expression as the cytoplasmic level of SSLENFRAYV peptide increases (represented on the x-axis). The other curves represent the SSLENFRAYV-H2Db surface expression in the presence of different ASNENMETM competitor concentration, down to the light blue bottom curve corresponding to gates P(1, 8) to P(8, 8) with the maximum level of competitor in untreated wild-type cells (top panel) or in IFN-γ treated cells (bottom panel). (D) Corresponding ASNENMETM-H2Db surface expression.

cleaved by cytoplasmic ubiquitin hydrolases, the idea being that the peptides are released at an equimolecular ratio to the fluorescent proteins, which can then be quantified using flow cytometry techniques, as described in Neijssen et al. 2005[95]. The released peptides can then be transported in to the ER and bind to MHC Class I molecules, and subsequently transit to the cell surface. A chimeric 1C3 monoclonal antibody was used to detect SSL-H2Db complexes on the cell surface, and an E10 Fab antibody was used to detect ASN-H2Db, and the abundance was then quantified using flow cytometry. The cytoplasmic concentration of the flourescently tagged proteins was varied from low to high for both SSL and ASN, and the peptide-MHC cell surface abundance was measured as these concentrations changed (Figure 3.1B). The collected data showed that the cell surface presentation of the more stable peptide-MHC complex SSL-H2Db was much higher than the ASN-H2Db complex for similar cytoplasmic protein concentrations, but that as the concentration of ASN increased in the cytoplasm for constant levels of cytoplasmic SSL, the cell surface abundance of SSL-H2Db decreased as ASN-H2Db increased (data points P(.,1 to 6) Figures 3.1 C & D top). An equivalent set of experiments was carried out but with mouse fibroblasts treated with mouse IFN-γ for 48 hours (data points P(.,1:6) Figure 3.1 C & D bottom), to determine the impact of IFN-γ on peptide competition. In general the presence of IFN-γ resulted in higher cell surface presentation of both SSL and ASN and reduced the competition between the peptides, as can be seen by the reduction in the spread of the SSL data points (Figures 3.1 C & D bottom), compared to without IFN-γ (Figures 3.1 C & D top).

3.3 A Mechanistic Model of Peptide Competition

The peptide filtering model[22] describes the supply of peptide from the cytoplasm in to the ER with the supply rate g_i, as described by Equation 2.36. Therefore, in order to use the model to predict peptide competition we needed to augment the peptide filtering model with several new factors (see Figure 3.2):

1. Conversion of intracellular abundance fluorescence to rate of peptide supply

2. Conversion of simulated cell surface copy number to antibody-based detection of cell surface pMHC abundance

3. The effect of IFN-γ on MHC-I and tapasin expression

4. Self/endogenous peptides

5. Peptide-specific on-rates

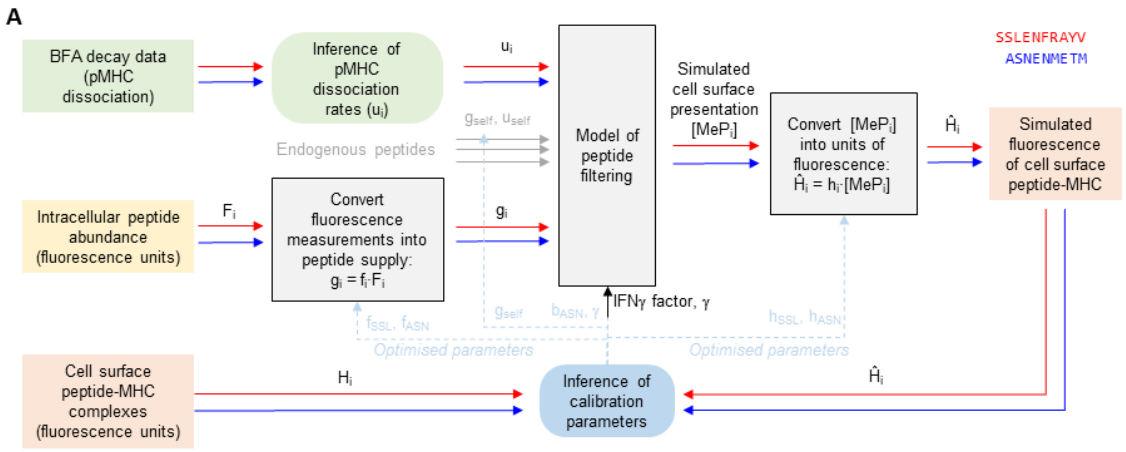


Figure 3.2: Diagrammatic representation of how we calibrated the model. The intracellular concentration of each peptide and their measured MHC unbinding rate were used as inputs to the model. We then fit the model to fluorescence data of the pMHC surface abundance. Square coloured boxes indicate measured or simulation data, grey boxes indicate models and rounded boxes represent inference algorithms. The red and blue line connectors represent peptide-specific information, and the dashed blue lines indicate that inferred parameters are eventually substituted back into the models for simulation/prediction.

To convert the intracellular abundance measured via Venus/mCherry fluorescence to ER supply rates for each peptide, we used two conversion factors f_{SSL} and f_{ASN} for SSL and ASN respectively, accounting for conversion from fluorescence units to numbers of molecules, intracellular degradation and peptide supply to the ER via TAP. As SSL and ASN were tagged with different fluorescent molecules, and will likely have different degradation rates and TAP affinities (and so different supply rates), the value of the scale factor will be different for the two peptides. Therefore, we define the supply rate g_i of each peptide i as:

$$g_i = f_i \times F_i, \tag{3.1}$$

where F_i is the fluorescent measurement corresponding to the intracellular abundance of peptide i.

A scale factor is also required to convert the fluorescent measurements of cell surface peptide-MHC H_i to number of complexes $[MeP_i]$, in other words the model output, denoted h_{SSL}, and h_{ASN} and this scale factor will be different for each peptide as different antibodies were used to detect them:

$$H_i = h_i \times [MeP_i] \tag{3.2}$$

We fit the model to two repeated SSL-ASN competition experiments, and as a result inter-experiment variations were observed in the measurement of SSLENFRAYV-H2Db at the cell surface using the 1C3 monoclonal antibody. Therefore, to obtain a better fit, the value of h_{SSL} was allowed to differ between the two experiments, $h_{SSL,1}$ and $h_{SSL,2}$ for experiments 1 and 2 respectively. To account for the addition of IFN-γ an up-regulation factor was required to increase the supply of MHC and tapasin. As with h_{SSL} we fit a different value of the IFN-γ up-regulation parameter for each of the two experiments that we will denote γ_1 and γ_2, to account for inter-experiment variations. The up-regulation factor was applied to both the MHC-I and tapasin supply rates, i.e. $g_M^{IFN} = \gamma g_M$ and $g_T^{IFN} = \gamma g_T$.

To achieve a better fit to the data we included a single additional peptide to represent the self-peptides present in the system, and the supply rate g_{self} was to be inferred. This single self-peptide was given an MHC unbinding rate of $u_{self} = 10^{-4}s^{-1}$ as this is the average peptide-MHC unbinding rate. The dissociation of ASN and SSL with the MHC-I molecule were measured using a Brefeldin A decay assay as described in Boulanger et al. 2009[18], and these measurements were used to determine the ASN and SSL MHC-I unbinding rates, u_{ASN} and u_{SSL} respectively. To further improve the fit to the data we also inferred the ASN binding rate to MHC-I, b_{ASN}.

We used Microsoft's Visual Genetic Engineering of Cells (GEC) software (freely available from `http://research.microsoft.com/gec`) to perform parameter inference. We specified the reaction system using the domain-specific Language for Biochemical Systems (LBS) and used an adaptive Metropolis-Hastings algorithm from the Filzbach software (`https://github.com/predictionmachines/Filzbach`) to perform the inference.

3.3.1 Parametrising the augmented peptide filtering model using Markov chain Monte Carlo

We wish to find the values of the set of parameters $\theta = \{f_{SSL}, f_{ASN}, h_{SSL,1}, h_{SSL,2}, h_{ASN}, \gamma_1, \gamma_2, g_{self}, b_{ASN}\}$ that best fits the data D. Assuming the model, H, we are using is correct, we therefore aim to approximate the posterior probability density $P(\theta|H,D)$, in other words the probability of the set of parameter values θ, given the data D. The posterior probability distribution is proportional to the product of a likelihood function $P(D|\theta,H)$ and a prior parameter probability density $P(\theta|H)$, according to Bayes' theorem:

$$P(\theta|H,D) = \frac{P(D|\theta,H)P(\theta|H)}{P(D)} \tag{3.3}$$

The likelihood function $P(D|\theta,H)$ describes the probability of the data D given the set of parameter values θ, and the prior $P(\theta|D)$ contains our beliefs about the distribution of the parameters beforehand. The normalising factor $P(D)$ is the probability of the data D over all possible values of the parameter set θ. This normalising factor cancels out during the inference process, and so does not need to be computed.

Under the assumption that the model H correctly describes the biological system, then any differences in the model output are due purely to experimental error, i.e. the data point y_i is normally distributed, where the mean of the distribution is the model prediction x_k at time $t = t_k$ and the variance σ^2 is, in this case, proportional to the measured fluorescence ($\sigma = \alpha\sqrt{y_k}$ for some α), therefore $y_k \sim N(x_k, \sigma^2)$. Therefore, the probability of each data point given parameters θ is given by:

$$P(y_k|\theta) = e^{-(y_k - x_k)^2/\alpha^2 y_k}/\alpha\sqrt{2\pi y_k} \tag{3.4}$$

Therefore, given the set of parameter values θ, the total likelihood function is the product of the probability of each data point y_k. When dealing with the product of very small numbers it is easier to compute the log-likelihood function, as follows:

$$l(\theta) := log(L(\theta)) = \sum_{k=1}^{N_d} log(P(y_k|\theta)) \tag{3.5}$$

For more details on MCMC see Robert & Casella (1999)[96]. Due to the complexity

Algorithm 2 Description of the steps in the Metropolis Hasting's Algorithm

1. Initialise the parameters at arbitrary starting value $\theta^0 = (\theta_1^0, \theta_2^0, ..., \theta_n^0)$
2. Draw candidate set of parameter values θ_{cand} from proposal distribution e.g. Gaussian distribution centered around θ^0
3. Evaluate the target distribution (here the log-likelihood) at θ_{cand}
4. Calculate log-acceptance ratio $r = l(\theta_{cand})/l(\theta_{current})$ and evaluate $\alpha = min[1, r]$
5. Generate random number $u \approx U[0,1]$
6. If $u \leq r$ accept proposed parameter values $\theta_{current} = \theta_{cand}$, else $\theta_{current} = \theta^0$
7. Iterate

of the model we cannot solve for the posterior probability density analytically, and so we must approximate it. We invoked the Filzbach software (available from `https://github.com/predictionmachines/Filzbach`) which uses a Metropolis-Hastings (MH) Markov chain Monte Carlo (MCMC) algorithm to sample from and approximate the posterior distribution. Monte Carlo sampling produces a sequence of values that have been drawn from some probability distribution. From these n samples, the expected value of the distribution can be computed. For very large values of n the distribution of the parameter values visited along the Markov chain converges to the true joint posterior distribution. Markov chains are stochastic models where the probability of an event occurring only depends upon the state of the system after the previous event. Therefore, in this case, the new values of the parameter vector θ_{t+1} only depends upon the previous parameter vector θ_t, and the probability of transition between these two states is described by the transition probability of the Markov chain. In the Metropolis-Hastings algorithm, a transition is accepted depending upon the ratio of the proposed transition to the likelihood of the current state. The system is initialised at arbitrary starting values for the parameters, then for each iteration a candidate set of parameter values θ_{cand} are sampled from the distribution, with a log-acceptance ratio, $r = l(\theta_{cand})/l(\theta_{current})$. If the ratio is greater than 1, and thus the likelihood is improved by the candidate parameter values, then they are accepted. If the ratio is less than 1, generate a random number u from the uniform distribution $U[0,1]$, and if $u \leq r$ accept the candidate parameter values,

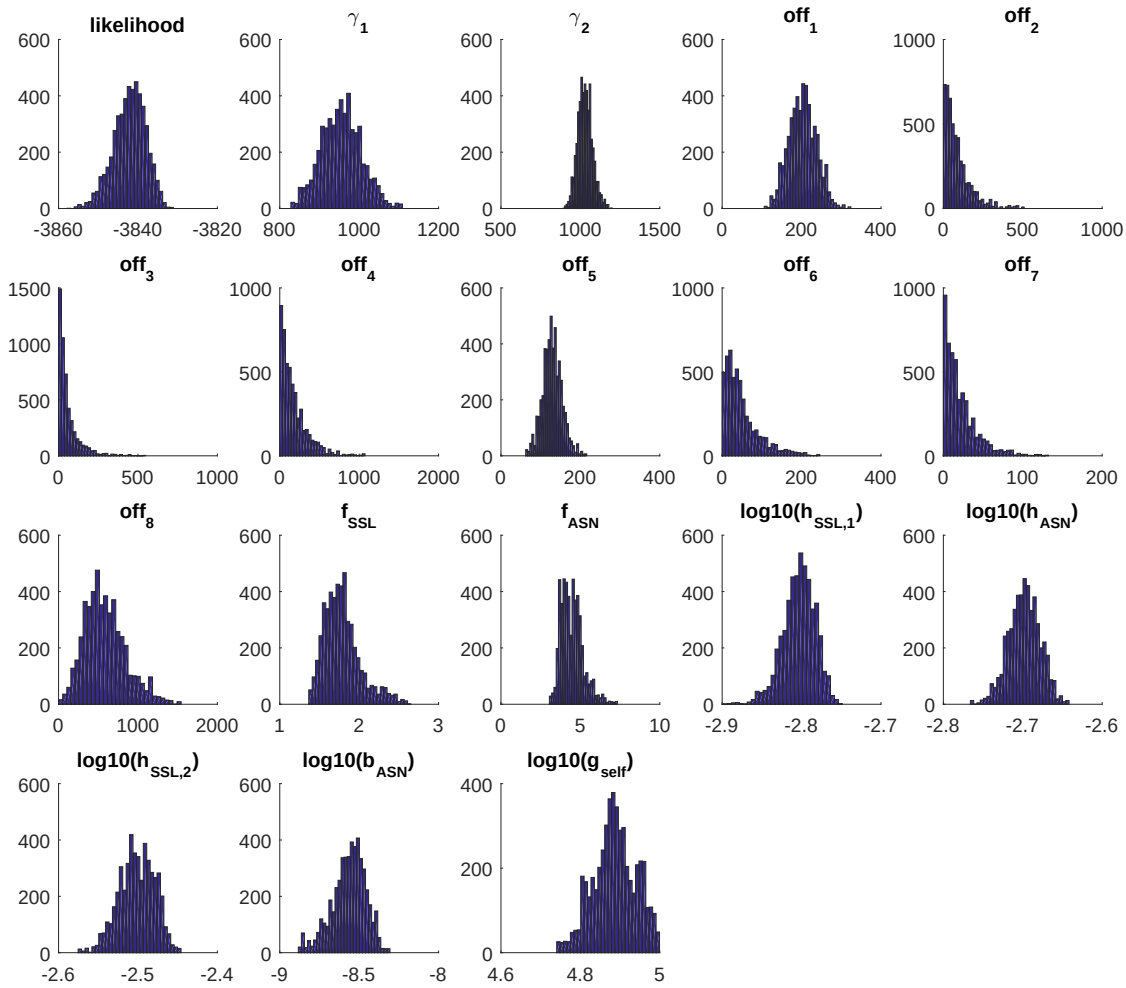


Figure 3.3: Marginal posterior distributions for calibration parameters. The marginal posteriors were established from 50,000 MCMC samples with a burn-in period of 20,000 runs. The parameter inference incorporated 2 experiments, for which γ and h_{SSL} had specific values (denoted by the superscripts 1 and 2). Each experiment measured SSL surface presentation or ASN surface presentation, and was either treated with IFN-γ or not (therefore, 4 scenarios per experiment). In all simulations shown in the manuscript, the parameter set that had the highest value of the log-likelihood function was used.

see Algorithm 2. To apply the Metropolis-Hastings algorithm to infer the parameters of our augmented model of peptide competition, we combined data from two separate experiments. We fit the model parameters to two sets of experimental data, where in each experiment the cytoplasmic concentration of SSL and ASN were varied and their cell surface abundances measured as described in Section 3.2. The cell surface abundance was measured in fluorescence and so each measurement will be associated with a background fluorescence. Therefore we were also required to fit for an additional eight parameters, $off_1, ..., off_8$, to account for the background fluorescence of SSL and ASN cell surface abundance with and without the addition of IFN-γ.

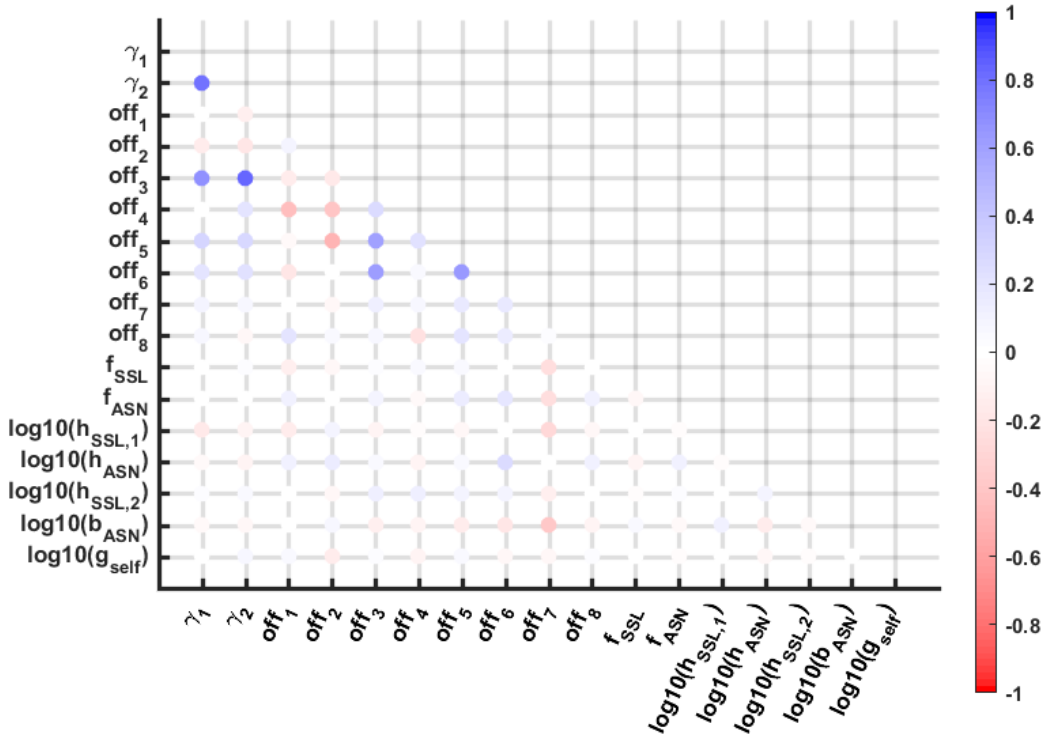


Figure 3.4: Pairwise correlations in the parameter joint posterior distribution. For each pair of parameters, the Pearson correlation coefficient was calculated over 50,000 MCMC samples with a 20,000 run burn in.

The MCMC algorithm was run for $50,000$ samples with a burn-in period of $20,000$ runs. Figure 3.3 shows the results of the MCMC fit, with the log-likelihood in the first panel and the marginal posteriors for the parameter values in the remaining panels. For each parameter, the value that corresponds to the highest value of the log-likelihood was

chosen as the best-fit parameter. From the marginal posteriors we can see that the majority of the parameters are approximately Gaussian distributed, except for the background offsets, with the exception of off_1 and off_8. The distribution of the offset samples are peaked at very low values close to 0 or in some cases lower than 0. This suggests a level of inaccuracy in the model that has resulted in lower-than-expected values of the background fluorescence being selected in order to try to compensate for this. Indeed, of the pair-wise correlations observed between the parameters (Figure 3.4), the largest magnitude correlations are between pairs of offsets, or offset-γ pairs. There is low pair-wise correlation between the rest of the parameters, where the colour bar ranges from 1 (full positive correlation) to -1 (full negative correlation). This suggests that the data is adequate to obtain well constrained best fit values for all parameters.

3.4 Results: Model Calibration

We fit the model to the data (Figures 3.5 and 3.6) as described in Section 3.3 and determined the goodness of fit by calculating the normalised root mean square error (NRMSE), where a lower value, closer to 0, means a better fit. The dots (experimental data) and lines (model simulation) are coloured to indicate the level of competitor peptide, so that the impact of peptide competition on the relationship between intracellular abundance and cell surface presentation can be compared conveniently. The highest SSL cell surface abundance $P(.,1)$ Figures 3.5 and 3.6 panels A and C corresponds to the lowest cytoplasmic abundance of ASN. However, in the presence of IFN-γ when we reach very high SSL cytoplasmic abundance the SSL surface abundance is no longer ordered according to ASN concentration. This suggests that at such high levels of cytoplasmic SSL and abundance of free MHC-I proteins to bind to, the competition due to ASN, when the ASN abundance is small i.e. $P(.,1) - P(.,4)$, has very little impact on SSL cell surface abundance. The model using the best fit parameter values fit well to the SSL surface expression without IFN-γ for both experiments, but the fit to SSL with IFN-γ has a lower NRMSE in the second experiment (Figure 3.6) than the first (Figure 3.5) with NRMSE of 0.0483 and 0.1284 respectively. We believe the reason for this is that for the first experiment, the model is overestimating the competition due to ASN in the presence of IFN-γ.

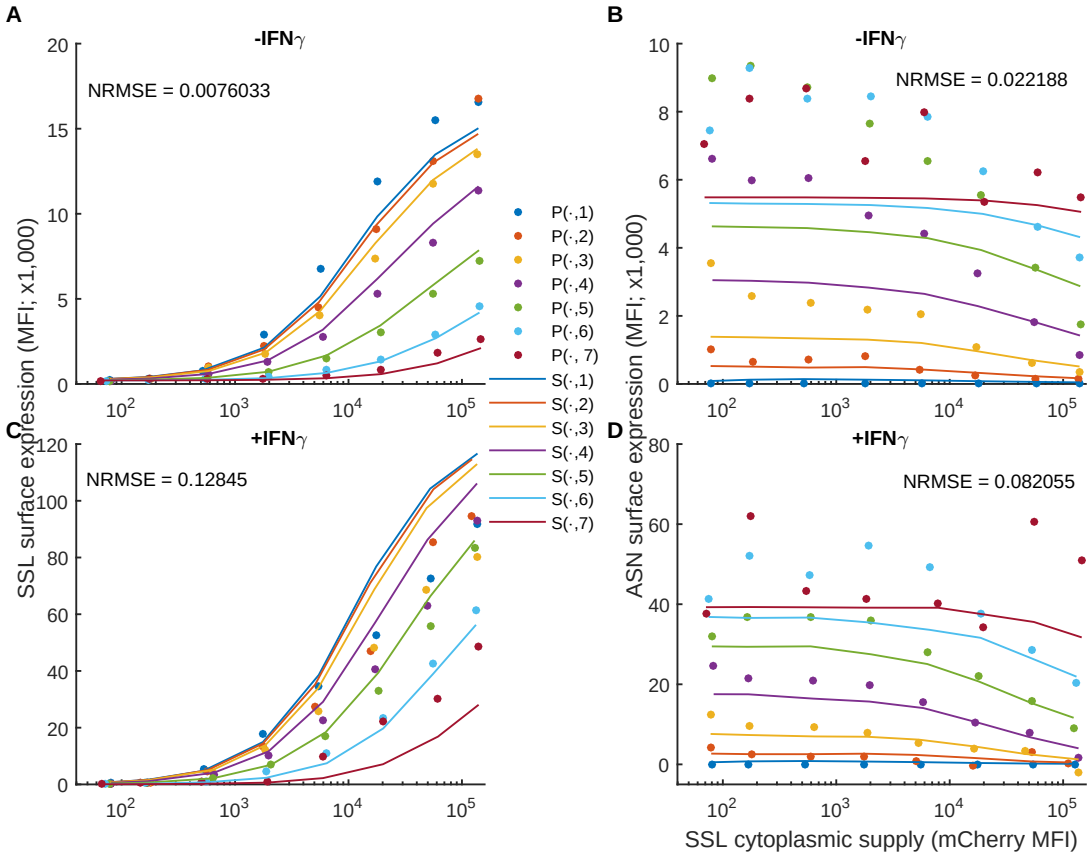


Figure 3.5: Model calibration, experiment 1 Comparison of the model (solid lines) evaluated at the maximum likelihood parameters values against experimental measurements (dots) for a single experiment measuring SSLENFRAYV vs ASNENMETM competition. The different colours represent different cytoplasmic levels of ASNENMETM, with P(.,1) and S(.,1) as indicated by the blue dots and lines respectively, representing the lowest ASNENMETM abundance, and P(.,7) and S(.,7) with the red dots and lines, representing the highest ASNENMETM abundance. The normalised root mean square error (NRMSE) between the data and the simulation is also indicated for each comparison.

The model is however able to capture the increase in both SSL and ASN cell surface abundance with the addition of IFN-γ (Figures 3.5 and 3.6 panels C and D), and Figure 3.6 C shows especially that the model is able to capture the narrowing in the vertical variation between the SSL cell surface levels for different constant ASN cytoplasmic concentrations. The averaged fluorescent measurements for cells with low levels of E10 fluorescence displayed trends that were inconsistent with the trends of intermediate and high expression levels. We attributed this to a low signal-to-noise ratio, and therefore suggest that these data are unreliable (Figures 3.5 and 3.6 panels B and D) and are not captured by the model as they are not a result of the competition between the peptides.

In spite of this the model is able to capture the general trend of the ASN cell surface abundance decreasing as the SSL cytoplasmic concentration increases (x-axis), whilst the ASN concentration stays constant (the individual coloured lines).

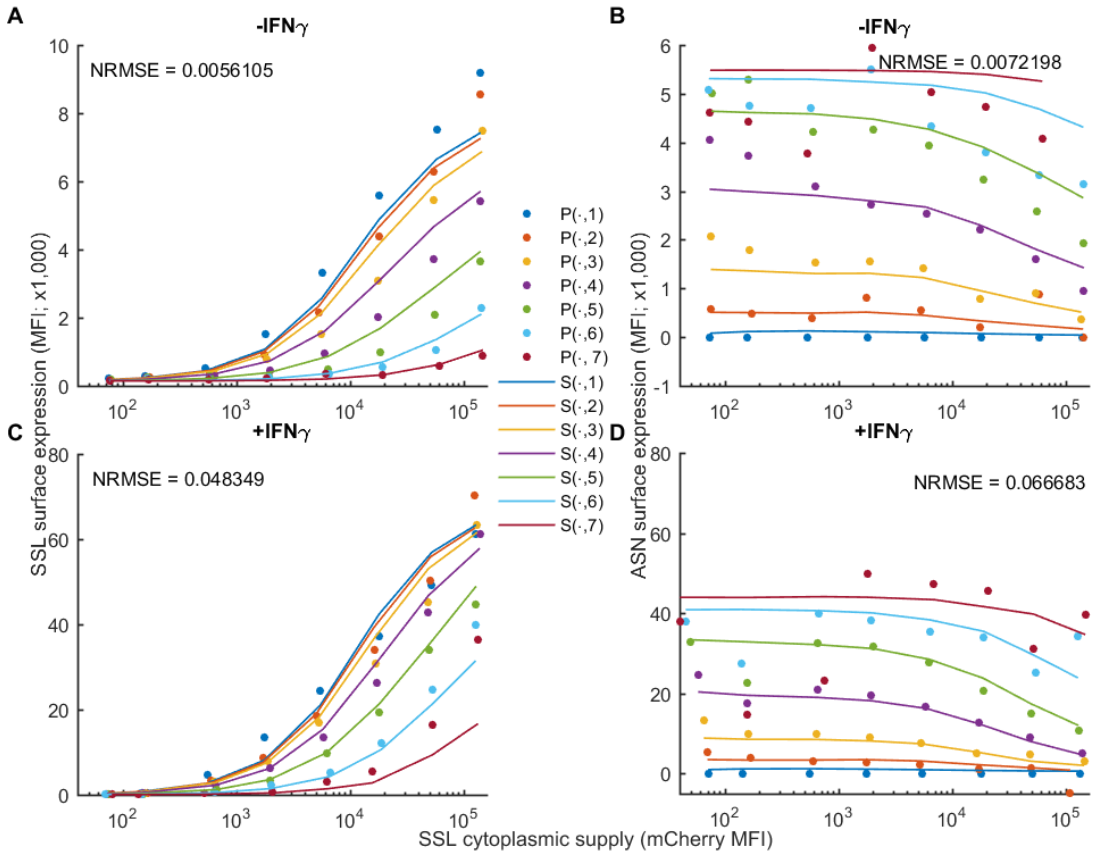


Figure 3.6: Experiment 2: Comparison of the model (solid lines) evaluated at the maximum likelihood parameters values against experimental measurements (circles) for a single experiment measuring SSLENFRAYV vs ASNENMETM competition. The different colours represent different cytoplasmic levels of ASNENMETM, with P(.,1) and S(.,1) as indicated by the blue dots and lines respectively, representing the lowest ASNENMETM abundance, and P(.,7) and S(.,7) with the red dots and lines, representing the highest ASNENMETM abundance. The normalised root mean square error (NRMSE) between the data and the simulation is also indicated for each comparison.

3.5 Results: The Calibrated Model Predicts Peptide Competition Well

In this study we wanted to know how the competition between two peptides is influenced by changing the off-rate of the competitor peptide, when the target peptide is in competition with varying competitor abundance. In turn, we wished to know how well our

augmented peptide filtering model, adapted from [22], was able to predict these interdependences. To this aim we calibrated the augmented model to flow cytometry data of

Peptide	**BIMAS score**	**NetMHC4.0 (nM)**	$t_{1/2}$ **min**	u_i **s**$^{-1}$	**E10 Score**
ASNE**A**METM	3.4	2208.8	43	2.7×10^{-4}	8.2
ASNENMET**A**	17	94.6	52	2.2×10^{-4}	14.2
ASNENMET**V**	17	13.2	132	8.8×10^{-5}	16.8
ASNENMET**L**	343	9.6	191	6.1×10^{-5}	11.8
ASNENMET**I**	343	6.6	212	5.4×10^{-5}	13.1
ASNENMETM	343	7.3	223	5.2×10^{-5}	11.5
ASNEN**L**ETM	411	10	238	4.9×10^{-5}	4.1
SSLENFRAYV	0.5	23.4	408	2.8×10^{-5}	3.8
AS**I**ENMETM	1029	3	456	2.5×10^{-5}	8.9
AS**I**EN**L**ETM	1235	3.6	502	2.3×10^{-5}	1.8

Table 3.1: Peptide sequences and off-rates BIMAS Score was determined using BIMAS predictor which can be found at `http://www-bimas.cit.nih.gov/molbio/hla_bind/` and the NetMHC 4.0 tool was used to predict the IC50 of H-2Db binding (nM). Peptide half-lives ($t_{1/2}$) were determined experimentally by BFA decay assays and converted into off-rates as log(2)/half-life (sec). The E10 score represents the MFI/1000.

peptide competition between a target peptide SSL and a competitor peptide ASN. Once calibrated, we then used the inferred parameter values in the model to predict the competition between SSL and several ASN variants, where amino acid substitutions in the sequence of ASNENMETM were chosen which change the unbinding rate of ASN with MHC-I. We used the BIMAS[61] tool to predict the peptide-MHC half-lives, and the MHC binding tool NetMHC4.0[62] to predict the peptide-MHC IC50 to chose the substitutions which would result in either a higher or lower off-rate than the original ASN sequence, and compared the two methods.

The half-lives ($t_{1/2}$) of the ASN variants in complex with MHC-I (H-2Db) were determined experimentally in brefeldin A decay assays [18], and comparison to the predicted values showed that in this case BIMAS performed better than NetMHC4.0 (Table 3.1). Our experimental collaborators performed another set of experiments, as described in Section 3.2, where the ASN variants were tagged with the fluorescent protein Venus and used in a competition assay against SSL tagged with mCherry, whilst the surface expression of the ASN variants was determined using an E10 Fab antibody. Only ASN variants that were recognised well by E10 were used. From these half-lives we were able

to compute the peptide-MHC unbinding rate as $u_i = log(2)/t_{1/2}$, and use these values when performing simulations to predict the competition between SSL and the ASN variants using the calibrated model parameters. Each ASN variant had a different binding affinity to E10. Therefore, to obtain a better fit, we scaled the predicted model ASN fluorescent intensity for each ASN variant by the ratio of their E10 affinity to that of ASN (see Table 3.1). We assumed that the point mutations have a limited impact on TAP affinity and so each ASN variant has a similar ER supply rate. The data shows a general trend of increased competition with SSL for H-2Db binding with increased half-life of binding of the variants (Figure 3.7 P(.,:), see Table 3.1 for pMHC half-lives), with the maximum competition observed in the presence of the slowest off-rate peptides ASIENMETM and ASIENLETM (off-rates of $2.5 \times 10^{-5} s^{-1}$ and $2.3 \times 10^{-5} s^{-1}$ respectively).

With the addition of IFN-γ, competition between SSL and the ASN variants was reduced in all cases, as can be seen by the smaller spread of data points (Figure 3.7, dots P(.,:)), and similarly for the slower off-rate ASN variants (Figure 3.8, dots P(.,:)). However, for the more unstable p-MHC complexes, the addition of IFN-γ was not sufficient to increase the cell surface level to recognisable levels, and there appears to be a low signal-to-noise ratio, as with the -IFN-γ case, as discussed earlier. As mentioned earlier we fit for two values of the parameter h_{SSL} to account for experimental differences between experiment 1 and 2. The best for value of $h_{SSL,1}$ for experiment 1 (Figure 3.5) was used in the predictions of SSL vs METV, whilst the value of $h_{SSL,2}$ for experiment 2 (Figure 3.6) was used in the predictions of the remaining peptides.

We determined how well the model was able to predict the competition by calculating the NRMSE between the simulation and the data. Figures 3.7 and 3.8 shows the data and the simulated SSL and ASN cell surface abundance respectively, against the data for SSL vs the ASN variants without and with IFN-γ, panels A and B on both figures, whilst panels C and D show the data and the simulated cell surface of the ASN variants. The model was able to predict the competition data very well for LETM vs SSL and METI vs SSL as demonstrated by the low NRMSE values, whilst the model predicted the competition between the remaining peptides reasonable well, as demonstrated by the slightly higher NRMSE values.

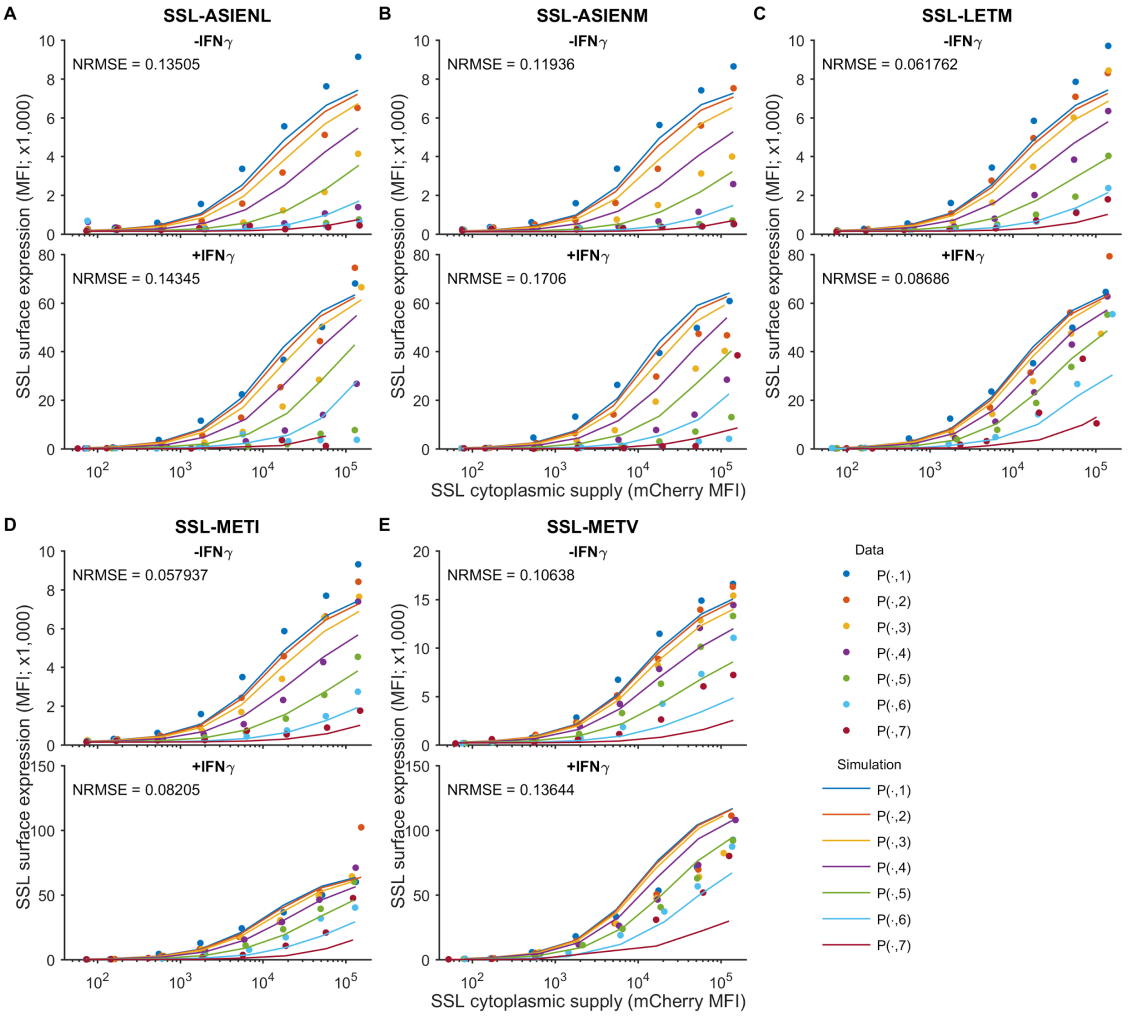


Figure 3.7: Predicting cell surface presentation of SSL when competing against ASN variants. The competition between SSL and ASN variants was predicted using the calibrated model. Each ASN variant had a different off-rate from MHC-I, as given in table 3.1. Each panel shows the predicted SSL presentation (coloured lines) against the fluorescent data (coloured dots) for SSL vs A) AS**I**EN**L**ETM, B) AS**I**ENMETM, C) ASNEN**L**ETM, D) ASNENMET**I** and E) ASNENMET**V**, both with and without IFN-γ. The normalised root mean-squared error (NRMSE) between the data and the simulation is also shown for each panel.

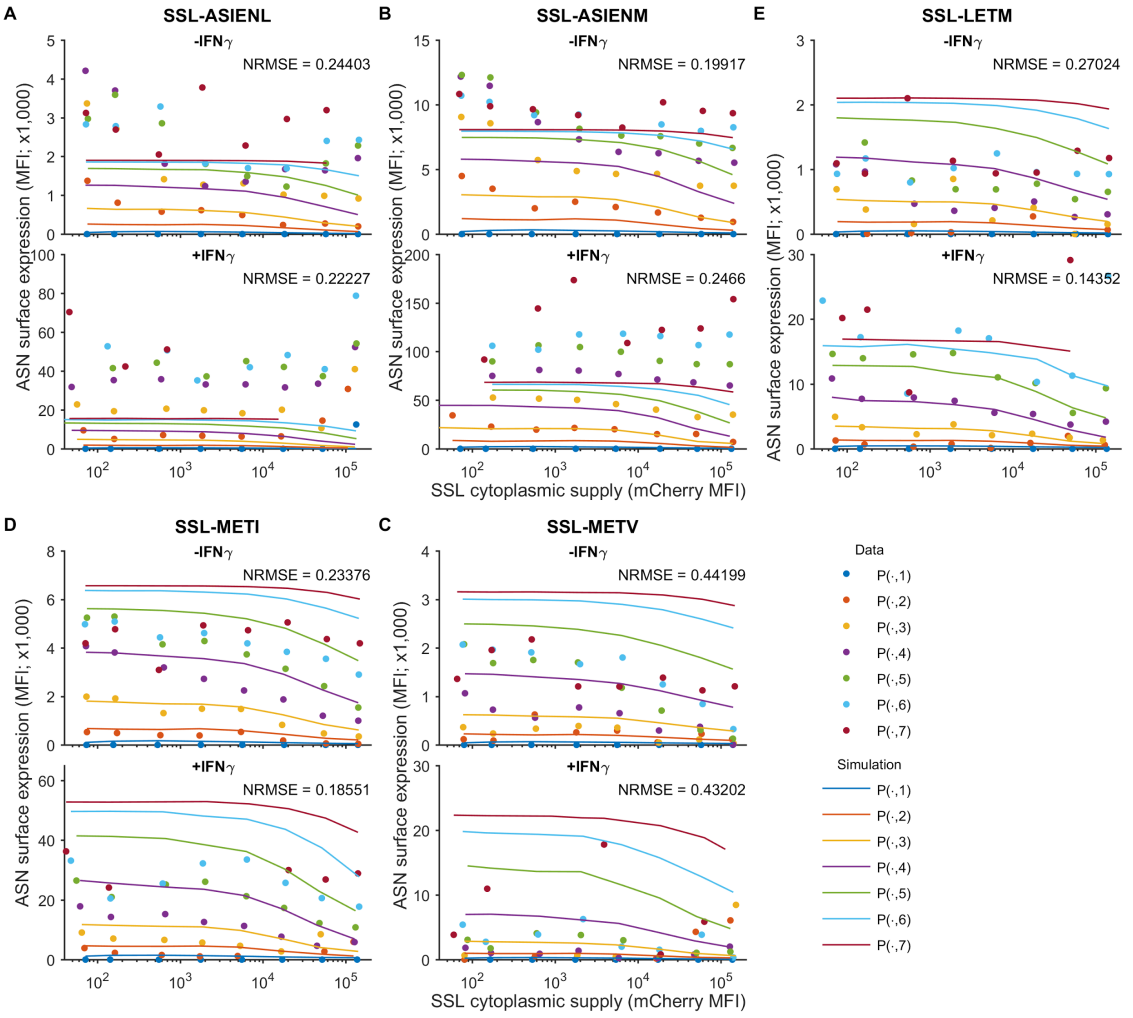


Figure 3.8: Predicting cell surface presentation of ASN variants competing against SSL. TThe competition between SSL and ASN variants was predicted using the calibrated model. Each ASN variant had a different off-rate from MHC-I, as given in table 3.1. Each panel shows the predicted ASN presentation (coloured lines) against the fluorescent data (coloured dots) for SSL vs A) ASIENLETM, B) ASIENMETM, C) ASNENLETM, D) ASNENMETI and E) ASNENMETV, both with and without IFN-γ. The normalised root mean-squared error (NRMSE) between the data and the simulation is also shown for each panel.

3.5.1 A simple peptide competition metric predicts cell surface abundance

We have demonstrated that by augmenting the peptide filtering model[22] as described in Section 3.3 (see Figure 3.2) we are able to predict the competition between two peptides of different off-rate and varying cytoplasmic concentration. However, we also wished to test if the simple peptide filter relation[22], in which the cell surface abundance of

a peptide is proportional to the ratio of its supply rate g_i to the square of its unbinding rate u_i^2, could provide a simpler and more intuitive method of predicting competition. This filter relation was used by Dalchau et al[22] to predict the equilibrium cell surface abundance of a pMHC complex.

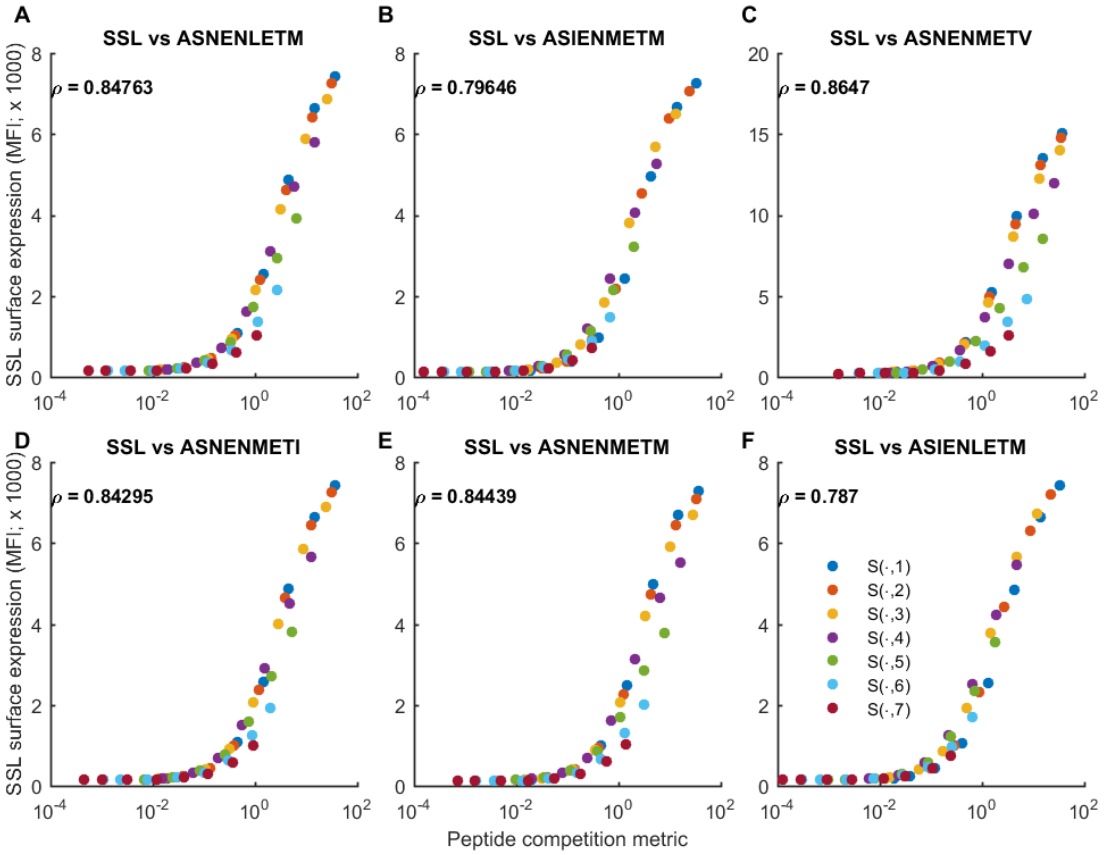


Figure 3.9: Cell surface abundance without IFN-γ can be predicted by a peptide competition metric. The peptide competition metric MeP_{ratio} (Eq. 3.6) was calculated using measurements of cytoplasmic peptide abundance for SSLENFRAYV and variants of ASNENMETM (horizontal axis) and compared with the simulated cell surface abundance of SSLENFRAYV (vertical axis), in the absence of IFN-γ.

The filter relation when used on its own does not provide a good estimation of the peptide cell surface abundance when that peptide is competing against other peptides, as it only approximates equilibrium cell surface presentation of a peptide in terms of its own supply and off-rate from MHC-I, and does not include the impact of competing peptides. Therefore, in order to account for the impact of the competitor peptide we normalised the filter relation. Initially we wanted to see how well the normalised filter relation could approximate the output of the peptide filtering model. In this case we are trying to predict

SSL cell surface abundance when in competition with ASN and self peptides, and we will approximate the supply rates of the two competing peptides g_i with the cytoplasmic peptide abundances. The calibration of the model included a self peptide supply rate g_{self} and unbinding rate u_{self}, and so we will use the best fit values of these parameters to represent the competition of the self peptides in the normalised filter relation. The normalised filter relation is therefore:

$$MeP_{ratio} \approx \frac{[SSL]_{cyt}/u_{SSL}^2}{([ASN]_{cyt}/u_{ASN}^2 + g_{self}/u_{self}^2)} \tag{3.6}$$

We calculate the normalised peptide filter metric for the SSL cell surface abundance for each competition experiment described above, with and without IFN-γ and calculated the Pearson's correlation coefficient between the metric and the corresponding model output. The different coloured traces represent the different cytoplasmic concentrations of ASN with which SSL is competing. There is a strong correlation between the normalised peptide filter metric and the model output for all experiments, both with and without IFN-γ (Figures 3.9 & 3.10 respectively), with all correlation coefficients greater than 0.7. The highest ASN concentration (red dots S(.,7)) result in the lowest SSL cell surface abundance, and correspondingly much lower values for the normalised filter metric. Similarly, the lowest ASN concentration (blue dots S(.,1)) results in the highest SSL surface abundance and also very high values of the metric.

If the metric is accounting for all important parameters that influence cell surface abundance we would expect similar values of the metric between traces to result in similar values of the cell surface abundance. This is indeed the case as can be seen as the traces in general follow the same trajectory as the value of the metric increases, showing a consistent relationship exists between the value of the metric and the cell surface abundance.

Once we had confirmed the metric was able to approximate the model, we wanted to see if it could be used alone to predict the experimental data. When applying the metric to the data, however, we do not have any data regarding the amount of self-peptide in the system. Using the calibrated values for g_{self} and u_{self} from the model resulted in a worse correlation (data not shown) than ignoring the self-peptide contribution entirely.

Therefore, the metric used to approximate the data was just the ratio of the peptide filter relation for the two competing peptides SSL and ASN:

$$MeP_{ratio} = \frac{f_{SSL}[SSL]^{MFI}_{cyt}/u^2_{SSL}}{f_{ASN}[ASN]^{MFI}_{cyt}/u^2_{ASN}} \tag{3.7}$$

To approximate the supply of each peptide - as we did in the model - we scaled the cytoplasmic MFI concentration of each peptide, $[SSL]^{MFI}_{cyt}$ and $[ASN]^{MFI}_{cyt}$ by the calibrated conversion scale factors f_{SSL} and f_{ASN} respectively.

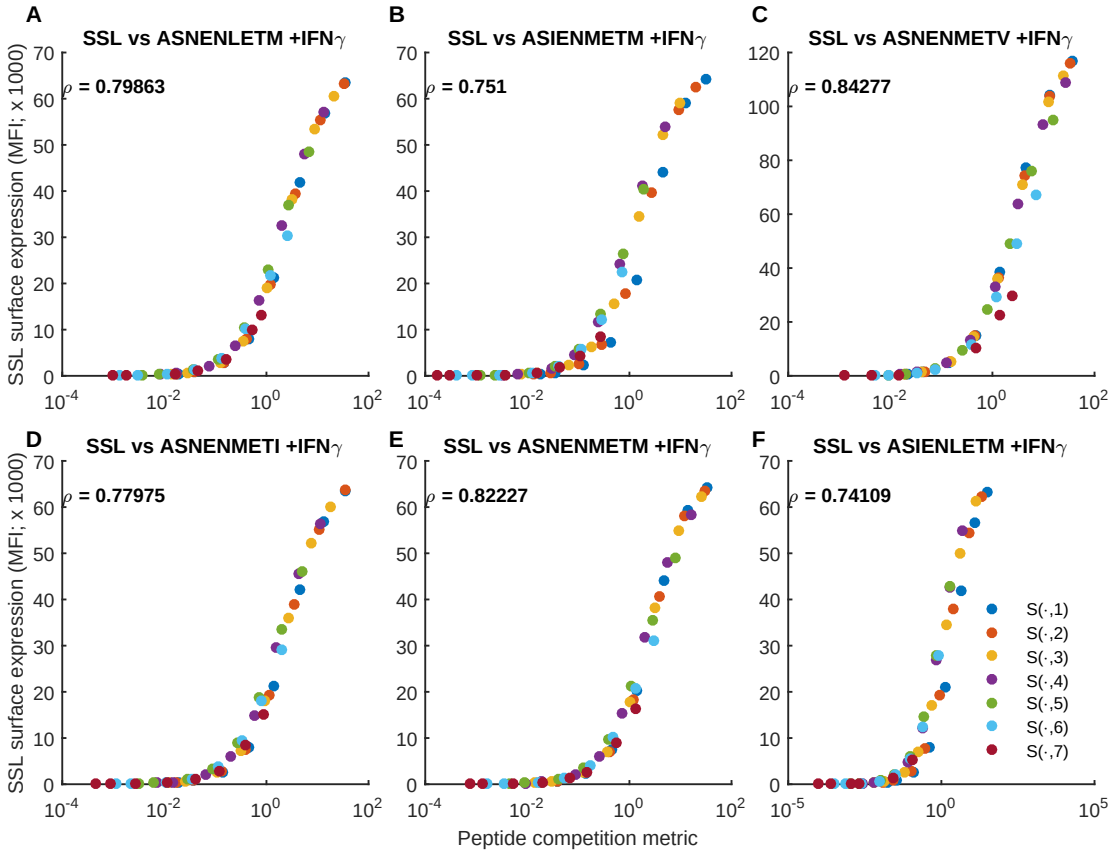


Figure 3.10: Cell surface abundance with IFN-γ can be predicted by a peptide competition metric. The peptide competition metric MeP_{ratio} (Eq. 3.6) was calculated using measurements of cytoplasmic peptide abundance for SSLENFRAYV and variants of ASNENMETM (horizontal axis) and compared with simulated cell surface abundance of SSLENFRAYV (vertical axis), in the presence of IFN-γ.

We observed a strong correlation between the metric and the data for SSL vs ASIENMETM and SSL vs ASIENLETM without IFN-γ (Figure 3.11 panels B and F respectively), whilst the remaining datasets had a weaker correlation with the metric with SSL

vs ASNENMETV having the weakest correlation of 0.52.

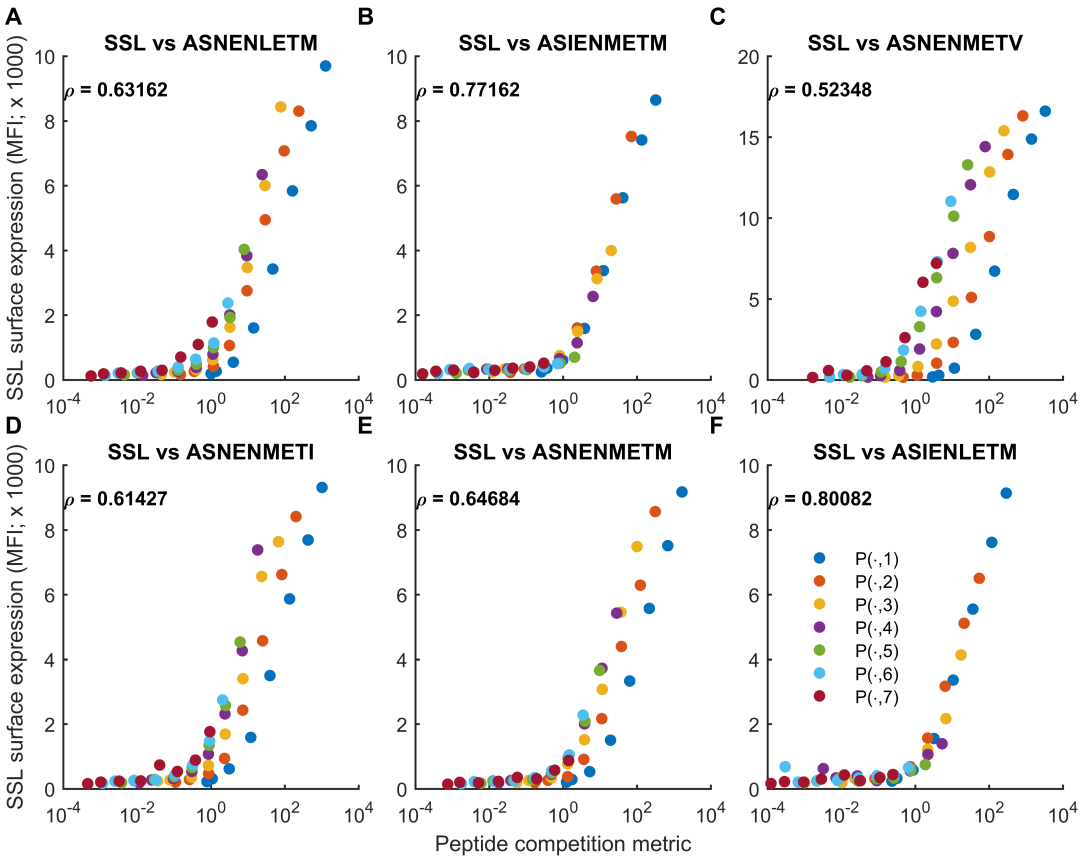


Figure 3.11: Cell surface abundance without IFN-γ can be predicted by a peptide competition metric. The peptide competition metric MeP_{ratio} (Eq. 3.7 was calculated using measurements of cytoplasmic peptide abundance for SSLENFRAYV and variants of ASNENMETM (horizontal axis) and compared with the experimentally measured surface abundance of SSLENFRAYV (vertical axis), in the absence of IFN-γ.

However, the metric was less successful at approximating the SSL cell surface abundance in the presence of IFN-γ, with all correlations less than 0.6 (Figure 3.12). As before if the metric were accounting for all important parameters influencing cell surface abundance, similar values of the metric should correspond to similar values of cell surface abundance. However we observed quite a spread in cell surface abundance for the same value of the metric between the traces for all experiments without IFN-γ with the exception of SSL vs ASIENMETM and SSL vs ASIENLETM, and similarly for +IFN-γ. This suggests the metric as applied here to the data is failing to account for an important parameter. The poor performance of the metric for the data compared to the simulations is most likely due to the missing contribution of the self-peptides in the data metric, and

so the metric is over-estimating the SSL cell surface presentation, especially when ASN cytoplasmic abundance is low e.g. P(.,1).

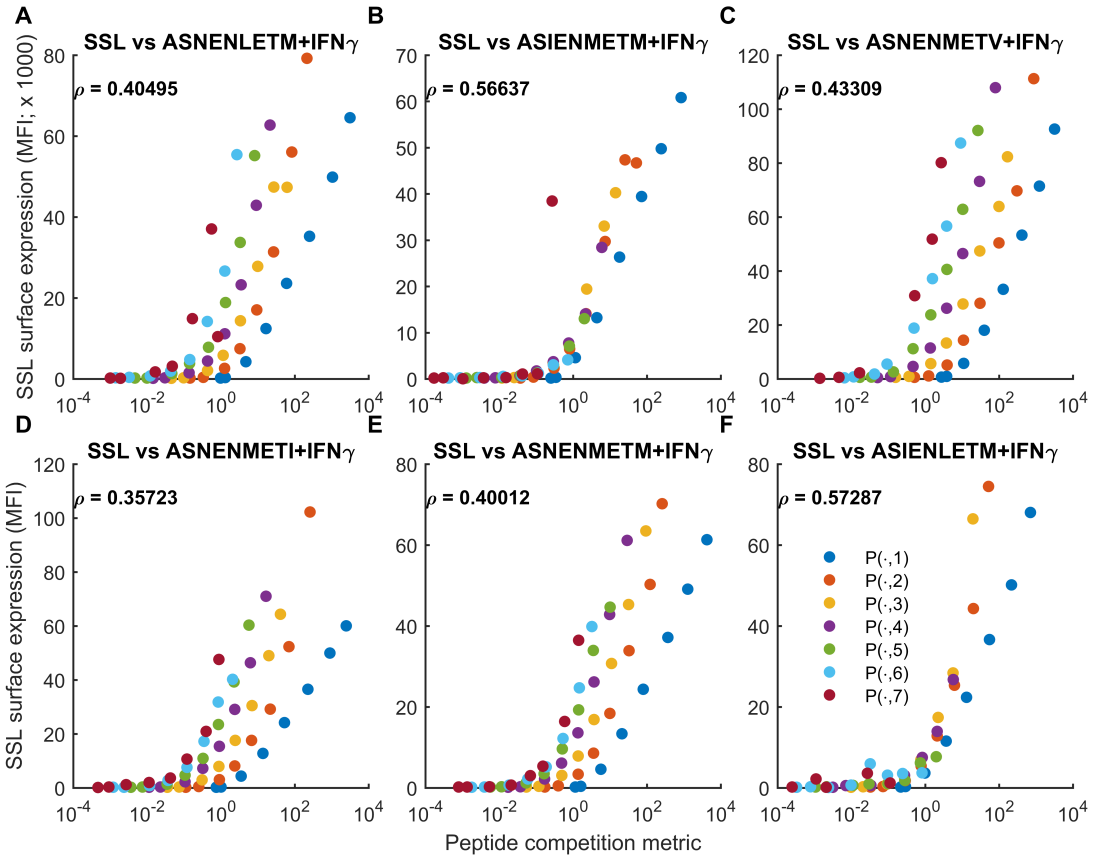


Figure 3.12: Cell surface abundance without IFN-γ can be predicted by a peptide competition metric. The peptide competition metric MeP_{ratio} (Eq. 3.7) was calculated using measurements of cytoplasmic peptide abundance for SSLENFRAYV and variants of ASNENMETM (horizontal axis) and compared with the experimentally measured cell surface abundance of SSLENFRAYV (vertical axis), in the presence of IFN-γ.

3.6 Discussion

The competition between peptides for binding and presentation by MHC-I in the ER is an important step in determining the T-cell response against cells infected with foreign pathogens, and understanding this step can aid in designing effective treatment and therapies for many infections and diseases. Dalchau et al. 2011[22] developed a mathematical model of antigen presentation by MHC-I, including only three main components: MHC, tapasin and peptide. The model was originally developed to explain how peptide optimisation by tapasin differs between MHC-I alleles. However, it was yet to be tested to see

if this model could predict peptide competition.

We augmented the peptide filtering model, so that the model could relate intracellular peptide abundance to cell surface abundance, via tapasin-assisted optimisation of pMHC loading and presentation. Using experimental data collected by our experimental collaborators, we calibrated the model and used it to predict the cell surface abundance of two competing peptides binding to the same MHC allele, where the off-rate of the competitor peptide varied. The calibrated model was able to capture the changes in cell surface abundance of a target peptide as the abundance of competitor peptide changes for competitor peptides with a range of unbinding rates.

The calibrated model was also able to account for the effect of the up-regulation of MHC and tapasin production following the production of IFN-γ, and the resulting decrease in the impact of increased abundance of the competitor pepitde on the target peptide. This agrees with Dudek et al. (2012)[97] where it was observed that presentation of the high affinity JAK-1355363 (SYFPEITHI), and the lower affinity IGRP2062 14 peptide (VYLKTNVFK) by H-2Kb, were both increased from $\sim 2,000$ to $\sim 15,000$ and from ~ 1 to ~ 25 copies per cell respectively after treating NIT-1 insulinoma cells with IFN-γ. A better understanding of IFN-γ is important as it is produced following viral infection and is associated with autoinflammatory and autoimmune diseases, and has been used in immunotherapy as an immunostimulant, however the mechanisms behind how it works are poorly understood. IFN-γ also has a role in cancer immunosurveillance[98] and tumour cells can also evolve immune escape mechanisms to block the IFN-γ pathway. Therefore, it would be beneficial to be able to predict CD8+ T-cell epitopes that are likely to be presented on tumour cells and the impact the addition of IFN-γ has on their subsequent presentation.

The increase in peptide presentation as a result of IFN-γ production may therefore alter which set of peptides are immunodominant. In our study the surface expression of the lower affinity peptide was enhanced in the presence of IFN-γ and Henrickson et al. (2008)[99] showed that T-cell activation can only occur above a threshold antigen dose. Therefore, the presence of IFN-γ could result in a CD8+ T-cell response to develop against a broader range of peptides.

We also found that a simple competition metric, Equations 3.6 & 3.7, based on the filter relation (Dalchau et al 2011[22]) correlated well with the model simulated cell surface abundance although the metric correlated less well with the corresponding experimental data, most likely due to the missing contribution of the self-peptides. The competition metric demonstrates the trade-off between the supply rate of the peptide - and thus its abundance in the ER - with the peptide-MHC unbinding rate on determining the cell surface abundance of a peptide. Such a simple metric could be used in vaccine design to quickly determine the concentration and pMHC half-life required for a peptide to be successfully presented in sufficient numbers on the cell surface, without the need for further experiment or simulation.

Chapter 4

A Mechanistic Model to Predict HeLa Cell Antigen Presentation

4.1 Introduction

On February 8^{th} 1951 cervical cancer cells were taken from **He**nrietta **La**cks (without her knowledge), a cervical cancer sufferer who later died from the disease. It was found that these cells could be developed, for the first time, in to an immortal cell line, meaning the cells would not die after a certain number of divisions, known as the Hayflick limit. These cells, known as HeLa cells, however, are cancer cells and so can multiply more rapidly than normal cells. As a result HeLa cells, have been used in a multitude of medical and biological experiments all over the world for many years, such as developing the polio vaccine and have furthered our understanding of cancer and are routinely used to test cancer treatments. HeLa cells are also used to grow viruses and further our understanding of infections such as HIV, and improve our understanding of the cell environment more generally. In 1999 it was observed that HeLa cells contain human papillomavirus (HPV) 18 DNA[100] and HPV18 has been linked to very aggressive cervical cancers. The HPV vaccine has been successful in preventing infection by specific types of HPV, including HPV16 and HPV18 which have are both associated with a high risk of cervical cancer. The HPV vaccine is a traditional protein vaccine, as it contains HPV L1 capsid proteins. This demonstrates how vaccines can be used to prevent cancer by providing immunity to the oncoviruses that cause them.

4.1.1 Cancer vaccines

Many cancers are not caused by viruses, and even in the case of HPV where a virus does cause cancer, low vaccination rates[101] mean that additional novel treatments are still required. Therefore a cancer vaccine has to either target an existing tumour (a therapeutic cancer vaccine), or prevent a tumour from developing (cancer immunoprevention). Indeed, the HPV vaccine would have no therapeutic effect on a person who had already developed cervical cancer as a result of HPV infection.

As an example of cancer immunoprevention, Nanni et al.[102] used an allogenic tumour cell vaccine (i.e. cancer cells taken from one patient which are then processed and turned in to a vaccine to prevent tumours of the same type) along side cytokines and other immunostimulants to prevent mammary carcinomas in a murine model.

Many therapeutic cancers vaccines are 'autologous', meaning they are made from tumour cell samples of a specific patient and are used to treat that patient, and so contain antigens specific to the individuals tumour[103].

Whole allogenic or autologous tumour cell vaccines have the potential to present the entire spectrum of tumour-associated antigens (TAAs) to the patient's immune system. However, one issue with autologous vaccines is that large samples of the patient's tumour are often required to make a vaccine that covers the entire spectrum of TAAs, which may not always be easy to acquire. Furthermore, tumours are heterogeneous, meaning that a vaccine developed to target a specific tumour may be ineffective against other tumours even within the same patient, as different antigens will be presented. Allogenic tumour cell vaccines, however, usually contain three established tumour cell lines, and their production is standardized and on a large scale, meaning clinical results are reliable and cost-effective[104]. However, the procedure is still complex and costly.

Peptide vaccines usually use several peptides of identified TAAs with the aim of stimulating a T-cell response in complex with MHC-I proteins. Advantages of peptide vaccines are that they are often more cost-effective and easier to produce than whole tumour cell vaccines, however, they cover a small spectrum of possible TAAs. A melanoma vaccine made up of six melanoma-associated peptides, demonstrated peptide vaccines to be safe and provided evidence of T-cell response to the melanoma[105]. As mentioned

in Section 1.2, strong protective T-cell responses have been observed against cancer epitopes, for example Toes et al.[25] reported such a response when vaccinating a murine model with epitopes from oncogenes required for tumour growth. However, at this point in time peptide vaccines are not optimized as they have a low response rate in clinical trials[106].

Therefore, a model that can predict epitopes derived from proteins expressed by oncogenes (mutated human genes involved in oncogenesis), may be able to help researchers more easily and quickly identify possible targets for T-cell vaccines for specific cancers.

HeLa cells have been used in an incredible number of experiments, and as a result there exists a large amount of readily available data regarding the HeLa cell proteome. In the previous chapter we demonstrated the augmented peptide filtering model can be used to predict the competition between a competitor and target peptide and that a simple metric provides a good approximation of the cell surface abundance of the target peptide as competitor abundance varies. We wanted to see if we could use this model to predict the presentation of an entire peptidome, and how well the filter relation correlates with the predicted cell surface abundance of each peptide. If the filter relation provides a good approximation to the predicted cell surface abundance of a peptide when competing against an entire peptidome, then it could be a useful tool with which to predict T-cell epitopes quickly and easily against a background of self peptides.

4.2 Methods

4.2.1 Obtaining quantitative measurements of the HeLa cell proteome

To be able to simulate the presentation of peptides derived from the HeLa cell peptidome we required a set of data measuring the abundance and half-life of the majority of proteins within the proteome of a cell. Boisvert et al.[107] quantified the intracellular abundance of a set of HeLa cell proteins using a method known as SILAC, in which an essential amino acid is supplemented with a non-radioactive, isotopically labelled form of that amino acid, which results in a mass-shift in the peptide compared to the control, or ‘light’ peptide.

The abundance of the protein is then measured as the intensity ratio of 'light' and 'heavy' peptides which correlates with the relative amount of the cognate protein from which the peptide derived. Nagaraj et al. 2011[108] measured the intracellular protein copy number of around 8000 HeLa cell proteins, and labelled each one using their International Protein Index (IPI) identifier.

In the first analysis we wanted to correlate the protein abundance, protein degradation, peptide-MHC off-rate and proteasomal cleavage probability with the peptide cell surface abundance predicted by the model. Therefore, we decided to use the Nagaraj et al. protein abundances, as they were in units of copy number, and therefore we would be able to predict an approximate peptide cell surface abundance, as opposed to the Boisvert et al. abundances which were quantified as an intensity ratio. We then compared the Nagaraj dataset to a dataset from Boisvert et al.[107] where the protein half-lives for 6000 HeLa cell proteins were measured in the cytoplasm, nucleus and nucleolus. We matched the Boisvert proteins with the Nagaraj et al.[108] proteins and found a cross-over of 4000 proteins. Out of these 4000 proteins only approximately 2000 had measured cytoplasmic degradation rates. We then converted the IPI identifiers to Uniprot IDs, from which we found the corresponding protein amino acid sequences from the Uniprot database(`http://www.uniprot.org/`). We then ran these sequences through IEDB MHC-I processing prediction tool[109]. This tool provides a predicted MHC-I binding affinity for each peptide, a proteasomal cleavage probability and a TAP affinity. We chose to run the predictions for all peptides 8-11 amino acids in length binding to HLA-A*68:02 as this MHC-I allele is known to be present in HeLa cells.

4.2.2 Calculating peptide-MHC off-rates

We calculated the unbinding rates using the approximation $u_i = b_P.IC50_i$ where $IC50_i$ is the IEDB predicted affinity of peptide i and b_P is the peptide-MHC association rate. The value of $b_P = 10^3 M^{-1} s^{-1}$ was chosen such that the off-rates of the most stable peptides were approximately $10^{-5} s^{-1}$.

4.2.3 Approximating proteasomal cleavage probabilities

Certain peptides will have a higher probability of being cleaved than others, because some proteasomes have preferences to cut peptides of a certain length, but also vary in their catalytic activity against different amino acids at the cut site. There are online machine learning tools which provide predictions for the likelihood that a given peptide sequence will be cleaved from the protein. The IEDB MHC-I processing tool[109] gives this value as a cleavage score, which is proportional to the logarithm of the amount of peptide generated from the cleavage of the peptides C-terminal. This score however, does not provide the actual probability of a peptide sequence being produced from the degradation of one protein. Experimentally measured values of proteasomal cleavage probabilities are sadly lacking in the literature, but it has been measured that the well known peptide SIINFEKL, or an N-terminally extended version of it, is produced via degradation of the OVA $6-8\%$ of the time it degrades[110]. SIINFEKL is known to be a highly immunogenic peptide, and so it is likely that it is produced with a high probability during proteasomal cleavage. Therefore, we set the highest possible probability of proteasomal cleavage to be 10%, and used this as an upper bound when predicting the proteasomal cleavage probabilities for the peptides in the model.

4.2.4 Model

To simulate the presentation of the HeLa cell peptidome from the protein data we had acquired, we used the peptide filtering model described in Section 2.6. To do this we had to provide an approximate supply rate for each peptide using the protein data and proteasomal cleavage probabilities we had acquired for each peptide. Therefore the supply rate of peptide i derived from protein j is:

$$g_i = A_{i,j} \cdot k_{i,j} \cdot ps_{i,j} \tag{4.1}$$

Here, $A_{i,j}$ refers to the abundance of the protein j from which peptide i is derived, $k_{i,j}$ refers to the degradation rate of the protein j from which peptide i is derived ($= ln(2)/\tau_{1/2}$ where $\tau_{1/2}$ is the protein half-life), and $ps_{i,j}$ refers to the cleavage probability of peptide i from protein j.

We then normalised the peptide supply g_i to ensure that when we summed the supply rate of each individual peptide it equalled the total supply of peptide from the original model which is important as the bimolecular rate constants (e.g. peptide-MHC association) were inferred at these levels.

The concentration of each peptide $[P_i]^*$ was assumed to be constant and at steady state. Therefore, in this case Equation 2.36 describing the peptide dynamics is equal to zero, and we simply write $[P_i]^* \approx g_i/d_P$, where d_P is the rate of peptide degradation. This approximation can be derived as follows.

Solving Equations 2.39-2.41 at steady state we get the following expressions for $[TMP_i]^*$, $[MP_i]^*$, and $[MeP_i]^*$:

$$[TMP_i]^* = \frac{(u_i+e)[MP_i]^* - b[M]^*[P_i]^*}{u_T v} = \frac{c[TM]^*[P_i]^*}{u_i q + u_T v} \tag{4.2}$$

$$[MP_i]^* = \frac{u_i[MeP_i]^*}{e} \tag{4.3}$$

$$[MeP_i]^* = \frac{1}{u_i}\frac{e}{u_i+e}(b[M]^* + \frac{x}{u_i+x}c[TM]^*) \overset{[TM]^* >> [M]^*}{\approx} \frac{1}{u_i}\frac{e}{u_i+e}\frac{x}{u_i+x}c[TM]^* \tag{4.4}$$

where $x = u_T v/q$ and we are assuming peptide loading takes place via the tapasin pathway and so $[TM]^* >> [M]^*$ Similarly, solving Equation 2.36 at steady state we obtain:

$$[P_i]^* = \frac{-u_i[MP_i]^* - g_i}{(c[TM]^*(\frac{u_i}{u_i+x} - 1) - d_P)} \tag{4.5}$$

If we substitute $[MeP_i]^* = e[MP_i]/u_i$ in to Equation 4.4 and rearrange for $c[TM]^*$ we get:

$$c[TM]^* \approx \frac{e[MP_i]^*}{[P_i]^*}\frac{u_i+e}{e}\frac{u_i+x}{x} \tag{4.6}$$

Substituting Equation 4.6 in to Equation 4.5 we arrive at the approximation:

$$[P_i]^* \approx \frac{g_i - e[MP_i]^*}{d_P} \tag{4.7}$$

Finally, if we assume high peptide turnover and thus rates of supply and degradation then $g_i >> e[MP_i]^*$, and therefore we can approximate $[P_i]^* \approx g_i/d_P$ (see [22] for more details). Therefore we assume that any peptide that egresses from the ER is rapidly replaced, ensuring the concentration of peptide in the ER remains at steady state levels throughout the simulation. This approximation is useful practically as it reduces the number of equations in the system and so makes simulating the model for a very large number of peptides more efficient.

4.2.5 Simulations

The simulations were carried out in MATLAB R2015b using the ode15s stiff solver to approximate the solution to the system of ODEs. We provided a Jacobian matrix to improve the efficiency of the solver and reduce the run time. We further reduced the number of peptides being simulated by removing all peptides where the unbinding rate was predicted to be higher than $u < 1 \times 10^{-2}/s$. Peptides with an off-rate higher than this limit will be unlikely to bind to an MHC-I allele long enough to be presented, therefore their impact on competition and the cell surface abundance of other peptides is likely to be very small. This lead to the simulation of 440,258 peptides in total. The supply rates to the ER were determined as described in Equation 4.1. The degradation rate of each protein was determined from the measured half-lives reported in the Boisvert et al.[107] dataset. The protein levels were kept constant at their measured copy number, and the peptide supply at steady state.

We analysed the output at $t = 10$ days to allow those proteins with long half-lives time to degrade sufficiently. We then chose only those peptides whose abundance at this time was predicted to be greater than or equal to 1. This is because a large number of peptides were predicted to have an abundance of less than 1 which does not make any physical sense and would skew the results of any analysis. The following analysis was therefore only applied to the 17,416 peptides with a cell surface abundance greater than or equal to 1 at $t = 10$ days.

4.2.6 Approximating cell surface abundance with the filter relation

We wanted to see how well the filter relation could approximate the simulation results in the case where a large number of peptides are competing for binding to MHC-I alleles. We calculated the filter relation for each peptide. The normalised filter relation, F_i, will approximate the relative cell surface abundance of each peptide compared to all others and for each peptide can be written as:

$$F_i = \frac{g_i/u_i^2}{\sum_{k,k\neq i}^{N} g_k/u_k^2} \tag{4.8}$$

where g_i is the supply rate of peptide i and u_i is the peptide-MHC unbinding rate, and g_k and u_k are the supply rates and unbinding rates of peptide k where $i \neq k$, and N is the total number of peptides.

4.2.7 Neo-epitope prediction

For the second analysis we wished to predict the presentation of novel epitopes (neo-epitopes) derived from mutated proteins present in tumours. Identifying immunogenic T-cell neo-epitopes is critically important for the development of cancer vaccines. Identifying possible neo-epitopes requires the identification of the corresponding mutated proteins. The Broad-Novartis Cancer Cell Line Encyclopedia (CCLE)[111] provides genetic mutation profiles for over 1000 cell lines, including HeLa cells.

Boegel et al. 2014[112] carried out an analysis where they predicted antigenic mutations for a large number of cancer cell lines including HeLa cells. They scanned cell-line specific mutations from both the CCLE and the COSMIC Cell Lines Project [113] and then used the IEDB resource platform to predict high affinity peptides that would bind to the HeLa cells endogenous HLA alleles, HLA-A*06:02 and HLA-B*15:03. However, in this analysis they did not incorporate the protein abundance, degradation or the proteasomal cleavage probabilities when predicting the mutant epitopes. Therefore, we decided to carry out an analysis where we predicted the presentation of mutant epitopes using our model of peptide presentation.

We chose to use the CCLE HeLa cell mutant dataset and searched for matches within

the Boisvert et al. and Nagaraj et al. HeLa cell proteome data. For this analysis there were a greater number of matches when using the Boisvert et al. abundances and half-lives. The HeLa cell CCLE mutant dataset identifies each protein using the Ensembl transcript (ENST) ID `http://www.ensembl.org/index.html`. Therefore we converted each protein's IPI from the Boisvert dataset to an ENST ID and matched the two datasets. We then obtained the protein sequences from the Ensembl database and inserted the amino acid mutations for each matched mutated protein from the HeLa cell dataset. We then simulated the presentation of peptides from all of the proteins in the Boisvert dataset that had both abundance and cytoplasmic half-life measurements, but this time including the mutated protein sequences. We used the intensity ratios for the abundance of each protein as we are only interested in relative levels of presentation in this analysis, as the data does not exist with which to test the predictions of the abundances of the mutated peptides. The half-lives, off-rates and proteasomal cleavage probabilities were determined as described in the previous sections. Once the simulation was complete we identified the peptides containing a mutation and normalised them by the most abundant mutated peptide on the cell surface, so that we could compare their presentation levels.

4.3 Results and Discussion

The cell surface abundance after 10 days of all peptides from all proteins is shown in Figure 4.1, where panels A, B, C and D show the simulated cell surface abundance plotted against peptide off-rate, protein half-life, protein abundance, and peptide cleavage probability respectively.

4.3.1 Correlation of cell surface abundance with individual parameters

We chose to look at the influence of these four components on peptide abundance as protein abundance, protein degradation, peptide off-rate and proteasomal cleavage have been reported to influence efficient HLA-I presentation[107, 91, 114, 22, 115, 116], and the probability of cleavage as this will determine the relative amounts of each peptide from each protein available in the cytoplasm. However, a weak correlation was observed between the peptide cell surface abundance and four parameters when considering each

parameter individually (Pearson correlations: protein abundance, $\rho = 0.031$; protein half-life, $\rho = -0.017$; peptide off-rate, $\rho = -0.0088$; and cleavage rate, $\rho = 0.0083$).

4.3.2 Correlation of cell surface abundance with normalised filter relation

We then calculated the filter relation (Equation 4.8) for each peptide (Figure 4.1 E) and found a very strong positive correlation ($\rho = 0.998$) with the model predicted cell surface abundance, demonstrating that the filter relation can be used to predict the relative cell surface abundance of an entire peptidome. The strong correlation between the filter relation and the cell surface abundance, compared to the weak correlations observed when considering the parameters on their own, suggests that efficient peptide-MHC presentation requires the optimisation of all four properties. For example, optimizing peptide-MHC affinity so that the unbinding rate is very slow (on the order of $10^{-5}s^{-1}$) does not guarantee a high cell surface abundance, as can be seen in Figure 4.1 C where we observe peptides with much higher off-rates having similar abundances to those peptides with much lower off-rates.

4.3.3 Correlation of cell surface abundance with raw filter relation

Whilst the correlation between the filter relation described by Equation 4.8 and the model predicted cell surface abundance is very high, it may not be always possible to calculate this value as it requires the characterisation of a large fraction of a cell's proteome. In cases where this data is not available, and a prediction of the relative abundances of peptides is required, the raw (un-normalised) value of the filter relation can be used, given as follows:

$$F_i = g_i/u_i^2 \tag{4.9}$$

To calculate this value an estimate of the supply of the peptide to the ER and a measure or prediction of the peptides MHC unbinding rate is required only for the peptide of interest, making it much simpler to acquire the relevant data. The correlation between the raw filter relation and the model predicted peptide cell surface abundance is high, with a Pearson's correlation coefficient of $\rho = 0.999$. Oddly, this is slightly higher than that of the normalised filter relation ($\rho = 0.998$). This is most likely because not one peptide

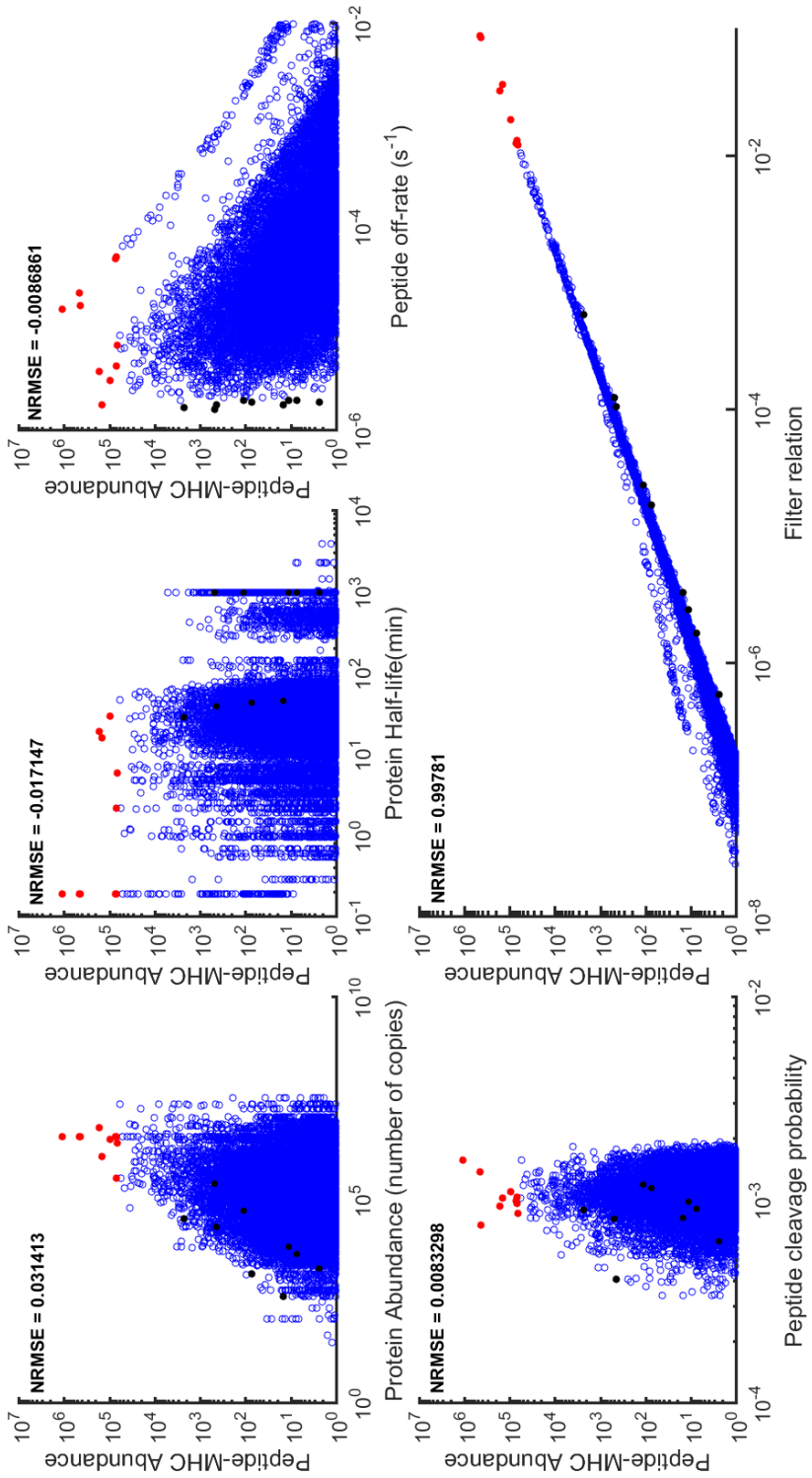


Figure 4.1: Simulating the HeLa cell peptidome The predicted abundance of peptides deriving from around 2000 HeLa cell proteins was simulated and the abundance at t=10days taken as the steady state. The correlation between the abundance and four important properties a) peptide off-rate, b) protein half-life, c) protein abundance and d) proteasomal cleavage was found to be low in each instance when considering the properties on their own. However, a strong positive correlation was observed between the cell surface abundance and the filter relation (panel e), which combines the four properties of protein abundance, degradation and peptide unbinding rate in to a single metric. This suggests a trade-off between the optimisation of all four of these parameters is required for efficient peptide-MHC presentation

dominates the competition in the ER and therefore there will be very little difference between the normalising factor for each peptides , i.e. $\sum_{k,k\neq i}^{N} g_k/u_k^2$ is approximately equal for each peptide, and so the normalising factor acts simply as a constant scale factor. This would not be true however if we were dealing with a case such as that presented in Chapter 3 where two peptides were competing in increasingly high abundance, suggesting they dominate in the cytoplasm and ER. If, however, a large number of peptides are competing and no single peptide or group dominates, each peptide faces the same level of competition from all other peptides. Therefore, the normalising constant is less important, and the cell surface abundance can be predicted using the raw filter relation given in Equation 4.9.

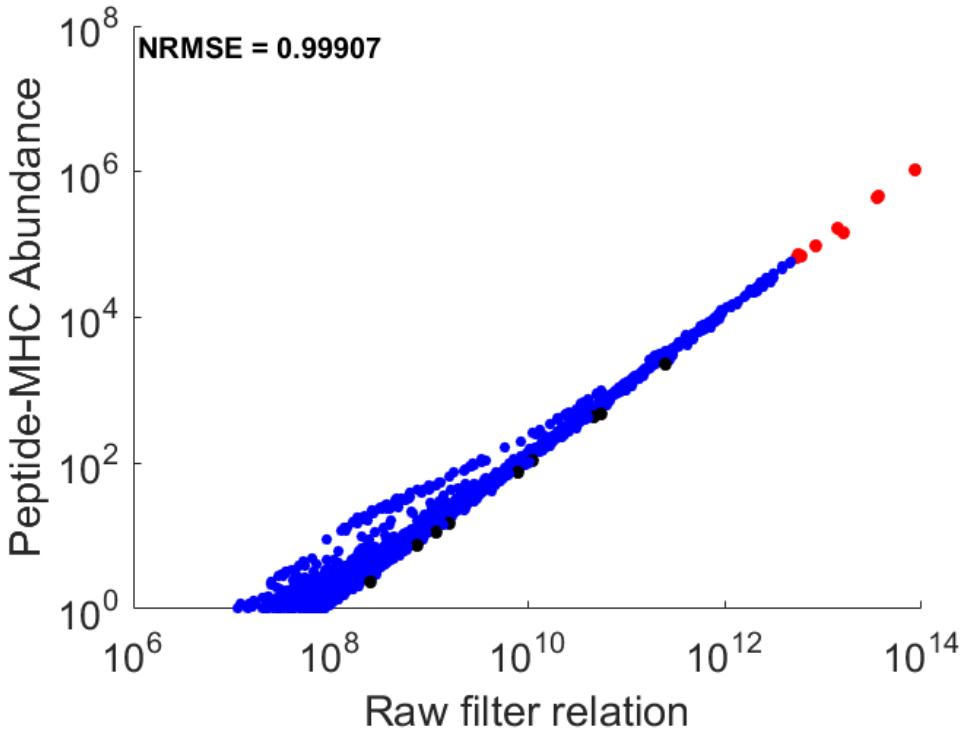


Figure 4.2: The relationship between the raw (unnormalised) filter relation and the cell surface abundance A strong positive correlation was observed between the cell surface abundance and the raw filter relation Equation 4.9, which combines the four properties of protein abundance, degradation and peptide unbinding rate in to a single metric, but does not normalise by the sum of this value for all other competing peptides. The correlation observed between the simulated cell surface abundance and the raw filter relation is higher than that of the normalised filter relation (Equation 4.8).

4.3.4 Comparison of predicted abundance with IEDB 'Total Score'

We wanted to see if the peptides that are predicted to have the highest affinity and highest cleavage probability are predicted to be the most abundant by the model. The IEDB MHC-I processing tool provides a 'Total Score' for each peptide, which combines the proteasomal cleavage score, the TAP affinity and the MHC-I affinity to predict a quantity that is proportional to the amount of peptide on the cell surface. The Total Score is the sum of the predicted TAP score, the proteasomal score and the MHC-I binding score. The MHC-I binding score is the -log10 of the IC50 value. In these simulations we have only used the proteasomal cleavage scores and the IC50 values predicted by the MHC-I processing tool. Therefore, we calculated an alternative 'Total Score' which is instead just the sum of the proteasomal cleavage score and the IC50 score. Those peptides with the top alternative 'Total Score' are identified as the black data points on Figure 4.1, whilst those peptides with the highest cell surface abundance as predicted by the model are identified as the red data points. As would be expected the peptides with the highest 'Total Score' all have very low unbinding rates (black data points on Figure 4.1 C), however, due to the wide range of cleavage probabilities and protein abundances and half-lives (black data points on Figure 4.1 D, A and B respectively) from which these peptides derive they do not correspond to the most abundant peptides as predicted by the model. The most abundant peptides predicted by the model all have low off-rates (though not the lowest), middling proteasomal cleavage probabilities, and derive from proteins with high abundances (though not the highest) and low to middling half-lives. This demonstrates the trade-off between these parameters in determining the cell surface abundance of a peptide, and this trade-off is captured in the filter relation. This highlights the importance of considering protein kinetics when predicting possible T-cell epitopes for use in cancer vaccines or immunotherapy.

4.3.5 The most abundant neo-epitopes

We predicted the relative cell surface abundance of neo-epitopes derived from HeLa cell proteins containing point mutations to predict which mutated peptides will be the most abundant (Figure 4.3). This simulation predicts that the peptides TVTGLTLLAV, GTFQNVSVQL and TVTGLTLLA, are the top three most abundant mutant peptides on the

cell surface. The set of mutant peptides predicted here as most abundant is entirely different from the set predict by Boegel et al. [112]. This may be because the overlap between the Boisvert et al. [107] data set and the CCLE[111] HeLa cell mutant dataset was quite small, meaning we may not have accounted for all of the important mutated sequences in this simulation. However, as we demonstrated in Section 4.3.4 the peptides with the highest affinity as predicted by IEDB do not correspond to the most abundant peptides on the cell surface, therefore the epitopes predicted by Boisvert et al. will most likely not be the most abundant mutant peptides on the cell surface.

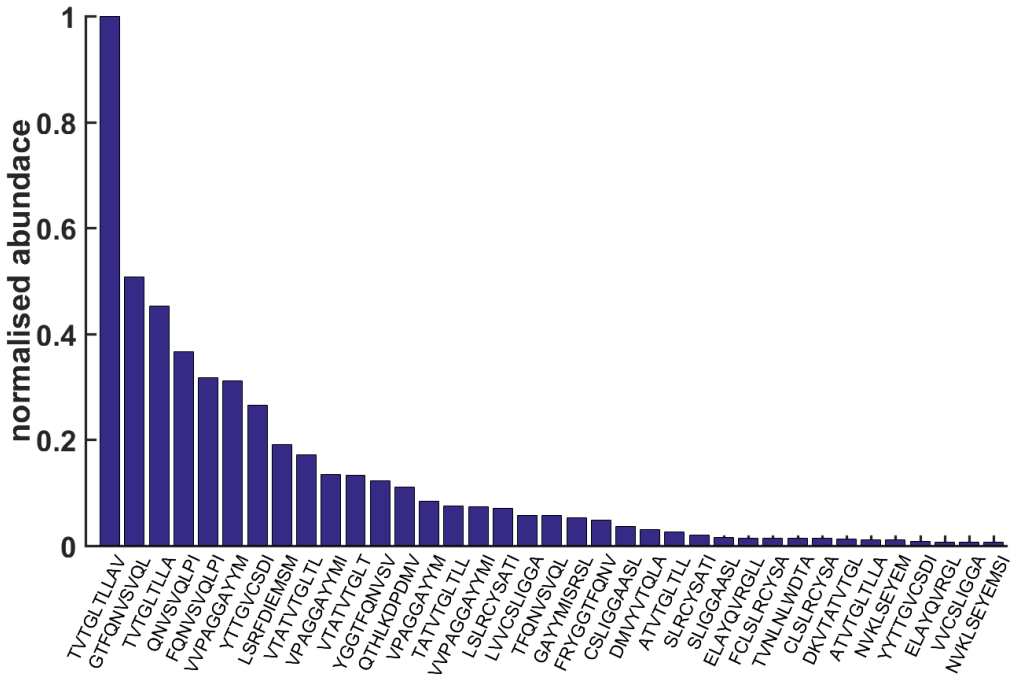


Figure 4.3: The relative abundance of the highest presented neo-epitopes, normalised to the abundance of the most abundance neo-epitope The top most abundant HeLa cell neo-epitopes predicted by inserting amino acid point mutations in to the protein sequences where matches between the Boisvert et al. [107] dataset and the Broad-Novartis Cancer Cell Line Encyclopedia (CCLE) HeLa cell mutations dataset.

4.4 Discussion

This chapter provides an idea of what predictions would look like for an entire peptidome. We have shown that the filter relation provides a better approximate to the simulated cell surface abundance of a peptide than protein abundance, protein half-life, peptide-MHC unbinding rate or proteasomal cleavage probability considered alone. We have also

demonstrated that when considering an entire proteome, all that is required to predict the cell surface presentation of a peptide is a measure of its supply - which here we approximated as the product of protein abundance, degradation and proteasomal cleavage probability - and the peptide-MHC unbinding rate. Unlike in Chapter 3, removing the normalising factor and using the raw filter relation provides as good a correlation as the normalised filter relation, likely due to the normalising factor being constant for each peptide as not one dominates.

Unfortunately, it is not currently possible to test the predictions of cell surface abundance of the HeLa peptidome, as the data are not available in the literature. Furthermore, at this moment, high-throughput methods for determining the cell surface abundance of a large number of peptides simultaneously do not exist. Indeed, the best attempt to date is that of Croft et al. 2013[21] where a mass-spectrometry method was used to quantify the presentation of eight vaccinia virus epitopes and the abundance of their source proteins simultaneously. The abundance of the source protein however was only quantified relative to its maximum expression, and not relative to the other proteins, therefore we could not use these measurements to test our model.

We have demonstrated that a model such as this could be used to predict the cell surface presentation of novel epitopes (neo-epitopes) derived from mutated proteins present in tumours. To improve the neo-epitope predictions we would require abundance and half-life measurements of the entire HeLa proteome, which can then be matched to the HeLa cell mutation datasets of CCLE and COSMIC. We hope that the development of new proteomics methodologies will enable such experiments to be carried out, paving the way for improved tools for the design of novel cancer treatments.

Chapter 5

A Mechanistic Model of Antigen Presentation Following Infection by Human Immunodeficiency Virus Type 1

5.1 Introduction

Human immunodeficiency virus type 1 (HIV-1) infects important immune cells such as T-cells and macrophages, resulting in low levels of CD4+ T-cells. These CD4+ T-cells are crucial to cell-mediated immunity, and so as their numbers decline, the infected person becomes more and more susceptible to opportunistic infections and tumours. HIV-1 infects a cell when the subunits gp120 and gp41 of the HIV protein Env, on the outside of the virion particle (as labelled in Figure 5.1), bind to the CD4+ T-cell receptors on the cell surface. The virion then fuses with the cell membrane and releases its contents in to the cytoplasm of the host cell. Once in the cytoplasm, reverse transcriptase transforms the single stranded RNA HIV-1 genome in to double stranded DNA, which is then imported in to the host cell nucleus and integrated in to the host genome via viral integrase.

The HIV-1 genome encodes for nine proteins: three viral structural proteins, Env, Gag and Pol, two regulatory genes Tat and Rev, and four accessory proteins, Vif, Vpu, Vpr and Nef. The host's cellular machinery then transcribes the HIV-1 genome to produce a full-length 9 kb mRNA. This full-length transcript encodes for the proteins Gag and GagPol, and is incorporated in to the budding virions. The full-length transcript can also

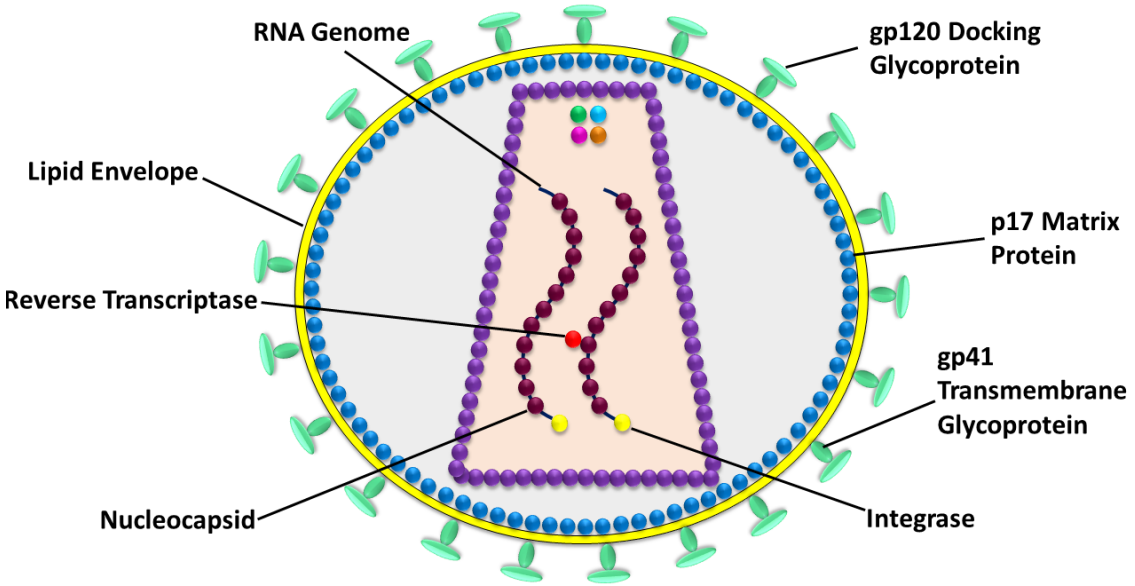


Figure 5.1: HIV-1 virion. Diagrammatic representation of HIV virion. The virion contains several copies of the viral proteins, as well as reverse transcriptase, integrase and the viral genome required for integration in to the host genome

be spliced once to produce 4 kb length mRNA which encodes for Env, Vpu, Vpr and Vif, or doubly spliced to produce 2 kb mRNA that encodes for Rev, Tat and Nef (Figure 5.2).

Only the doubly-spliced 2 kb mRNA is small enough to be independently exported from the nucleus in to the cytoplasm[117], where it is translated in to either Rev, Tat or Nef. Rev is then imported back in to the nucleus, where it binds with the Rev Response Elements (RRE) of the full-length and singly-spliced transcripts (see Figure 5.2). Once a threshold concentration of RRE-bound Rev is reached, the full-length and singly-spliced transcripts can then be exported to the cytoplasm.

Tat is also imported back into the nucleus where it binds with the trans-activating response element (TAR), and increases the rate of transcription up to 100-fold. Full-length mRNA in the cytoplasm is then translated to produce Gag and Pol, and singly-spliced mRNA is translated in to Env, Vpr, Vpu and Vif. The full-length mRNA and translated HIV-1 proteins then assemble at the cell membrane and create a new viral particle, which is then released from the cell.

There are three main stages of HIV infection. Primary (or acute) HIV infection, shortly following contraction of the virus, presents in many individuals as flu-like symp-

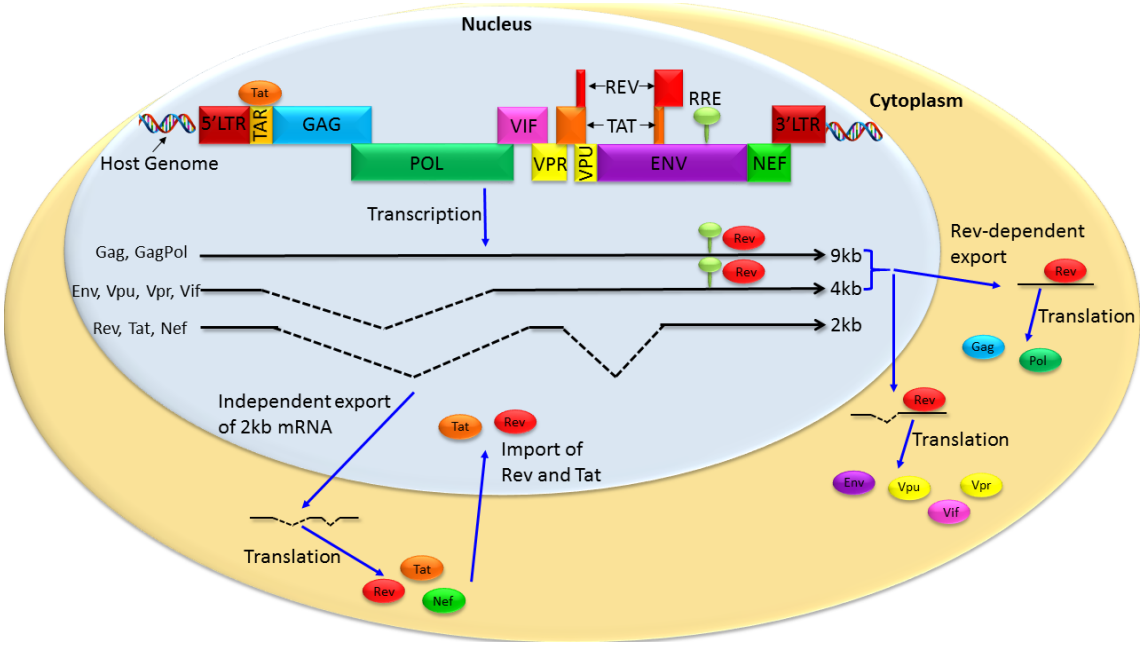


Figure 5.2: HIV-1 intracellular kinetics. Once the HIV genome has been integrated with the host DNA, the host machinery can be used to transcribe the HIV proteins. Full-length (9 kb) mRNA encodes for the structural proteins Gag and GagPol, 4 kb mRNA encodes for the Env, Vpu, Vif and Vpr, whilst the 2 kb mRNA encodes for the regulatory proteins Nef, Tat and Rev. The mRNA 9 kb and 4 kb mRNA requires Rev binding for export to the cytoplasm, whilst the 2 kb mRNA can export independently. Once in the cytoplasm the mRNA is translated using the host machinery in to the HIV proteins.

toms, which last between 2-4 weeks post infection, although many patients can be asymptomatic. This stage is associated with rapid viremia production and depletion of CD4+ T-cells, and an expansion in the number of CD8+ T-cells[118]. Cytotoxic T-cells, also known as CD8+ T-cells, express a T-cell receptor (TCR), which interacts with peptide-MHC class I complexes on the cell surface (see Section 1 for more detail). If this MHC-I molecule is presenting an HIV-1 peptide, then the CD8+ T-cell should recognise the presence of the virus and destroy the infected cell. However, in the majority of patients, this CD8+ T-cell response is not enough to control the spread of the virus inside the body.

Following primary infection, there is a phase of clinical latency, where the infected individual is largely asymptomatic, a period which can last from between 3 years up to over 20 years, depending on the individual[119]. A small subset of infected individuals maintain high CD4+ and CD8+ T-cell counts, along with low viral loads ($< 10,000$ HIV-RNA copies ml^{-1}), without receiving any anti-viral therapy or treatment, for up to 25 years[120]. Such individuals are known as long-term non-progressors (LTNP). Less than

1% of LTNPs, known as Elite Controllers (EC), have undetectable viral loads of < 50 HIV-RNA copies ml^{-1}[121][120] and exhibit very slow disease progression.

5.1.1 HLA and HIV-1 rates of progression

Understanding the factors that influence HIV progression rates and lead to long term control will aid in the design of vaccines and improved treatments. Long term non-progression is not linked to viral defects or polymorphisms[122], but has instead been linked to the response of the host immune system. For example, HIV can be transmitted from an individual experiencing fast progression to AIDS to an individual who then becomes an elite controller[121]. Several factors are known to influence how long this latency period lasts for, and so how long it takes for an individual to progress to acquired immunodeficiency syndrome (AIDS)[121]. One such factor is the expression of HLA alleles encoded for by the individuals genome. Several HLA alleles such as B*58, B*57, B*27 and B*14 have been found to be over-represented among LTNPs and ECs, and they are associated with Gag-specific CTL responses[123, 124, 125, 126], which originate from highly conserved regions of the Gag protein sequence[127]. For example, both B*57 and B*58 are associated with CD8+ T-cell responses against Gag epitopes TW10 (TSTLQEQIGW) and KF11 (KAFSPEVIPMF)[128]. The known T242N escape mutation in TW10 leads to diminished viral replication capacity[129], as does the A163G mutation of KF11[130].

Furthermore, escape mutations within these CTL epitopes are associated with the eventual transition of LTNPs and ECs to progressors[131]. There is also evidence which suggests that allele combinations, such as HLA-B*57:01-Cw0602 and HLA-B*27:05-Cw0102 have an even stronger effect on disease progression, indicating that the effect of control can be additive if an infected individual expresses several HLA alleles which are associated with LTNP[132].

Not only are HLA-B alleles associated with control of HIV, but there are also some alleles over-expressed among individuals who progress very quickly to AIDS, such as HLA-B*35 and -B*18. These non-controlling alleles are associated with CTL responses against non-Gag epitopes, such as Nef and Env epitopes[133], and mutations in these epitopes are fitness neutral[134]. The Env and Nef proteins are both highly variable, with

Env being the most variable sequence in the HIV genome[123] and so this may explain the inability of Env- and Nef- specific CD8+ T-cells responses to control the progression of HIV.

It is in fact Pol and not Gag that is the most conserved HIV protein[123], and so the strong association between Gag and control of HIV progression may be influenced by factors other than sequence conservation. Not only are these Gag epitopes highly conserved, but the Gag protein itself is also the most highly abundant in both the HIV virion and the host-cell cytoplasm during the replication cycle. There are around 4900 copies of Gag per HIV virion[135], with a Gag-Pol ratio of 20:1 per virion. Similarly, the ratio of Gag to GagPol synthesis from full-length mRNA also estimated to be around 20:1 [136]. In a study in to the Simian Immunodeficiency Virus (SIV, a close relative of HIV that infects Macaques), showed CD8+ T-cells against Gag-specific epitopes as early as 2 hours post infection, whilst non-Gag specific responses, such as those to Nef and Env, were not detected until *de novo* synthesis of HIV proteins occurs during the replication cycle, at around 10 hours post infection[137].

We therefore wish to determine how the large concentration of Gag in both the virion and during replication effect Gag peptide presentation compared to peptides from other proteins, and how this impact control of disease progression. To do this we first determined if the IEDB MHC-I processing tool could predict the dominance of Gag peptides, then we constructed a mechanistic model of HIV-1 intracellular kinetics and antigen presentation and simulated the system deterministically to determine if including Gag protein kinetics could shed some light on Gag epitope dominance. We also compared the HIV-1 peptide presentation between a group of controlling and non-controlling alleles to see if there was any discernible difference in the peptides presented by the two groups. Finally we constructed a mechanistic model of the presentation of HIV-1 virion derived peptides, which we simulated stochastically due to low molecular copy numbers, to determine the earliest time point at which we would expect to see antigen presentation post-infection of the cell.

5.2 Predicting HIV-1 Peptide Presentation using Existing Machine Learning Tools

We wanted to see if the Immune Epitope Database (IEDB) epitope prediction methods used alone could explain why Gag epitopes are so immunodominant and why their expression is correlated with long term non-progression. We used the IEDB MHC-I processing tool, which combines predictions of proteasomal cleavage, TAP transport and MHC class I binding[109] and outputs a 'Total Score' for each peptide, which is designed to be proportional to cell surface abundance of the peptide. Thus, a higher total score means the peptide has a better probability of being presented on the cell surface and so the more immunodominant the peptide.

5.2.1 Methods

We used the HIV-1 clade C consensus protein amino acid sequence from `http://www.hiv.lanl.gov` as the input to the IEDB MHC-I processing tool and compared the epitope predictions for each protein. We performed this analysis for four alleles associated with long term control of HIV: HLA-B*58:01, -B*57:01, -B*27:05 and -B*44:03 [123, 124, 125, 126, 138], and four alleles associated with fast progression: HLA-B*18:01, -B*35:03, -B*07:02 and -B*55:01[133, 139].

We used two different methods to determine which peptides out of the total predicted data set for each allele to consider as binding peptides. In the first method, we used a threshold of 500 nM for the MHC-I binding predictions[140](Figure 5.3 A & D), whereas in the second method we took the peptides within top 1% Total Score when all possible peptides for the HIV proteome were considered (Figure 5.4 A & D). We applied two different methods because whilst IEDB recommends a cut-off of 500 nM when predicting a binding epitope, the prediction scores for different HLA molecules cannot actually be directly compared. Therefore, we wanted to see if there was a discernible difference in the results when using the two different methods.

We also compared each HIV protein by the average Total Score of peptides from each protein (Figure 5.3 and 5.4 B & C), and recorded the total number of peptides from each protein that were predicted to be presented for each method. For both methods,

due to the strong association with the presentation of Gag epitopes, we would expect that for the alleles associated with long term non-progression, the highest Total Scores would come from Gag peptides and on average Gag peptides would have a higher Total Score.

Borghans et al.[141] used NetMHC3.0, a peptide-MHC binding prediction tool, available at `www.cbs.dtu.dk/services/NetMHC-3.0/` to predict the affinities of HIV peptide to different HLA alleles, and then ranked each peptide among all HIV peptides from highest to lowest affinity, where a lower rank corresponds to a high affinity binder. They compared the predicted ranks of peptides from different HIV-1 proteins between a group of controlling alleles (HLA-B*27:05, -B*57:01 and -B*58:01) and a group of non-controlling alleles (HLA-B*35:03 and -B*53:01). They then looked at only the top 3 ranking Gag epitopes and found the controlling group had a significantly higher preference for Gag than the non-controlling group. Doing the same for the other proteins they found the non-controlling group had a preference for Nef peptides over the controlling group. Significant differences in the preferences for Vpu, p17, Vif and Ref epitopes were also observed but it was concluded that the median ranks of these peptides were so high (i.e. the peptide affinities were so low) that the differences were most likely not physiologically important.

However, they only considered the top 3 peptides from each protein in their analysis, and did not include proteasomal or TAP predictions. Therefore, we carried out a similar analysis to determine if there was any statistical significance between the predicted Total Scores of the peptides from each HIV-1 protein in the top 1% of each allele, grouping by control vs non-control, using a Wilcoxon rank sum test.

5.2.2 Results: method 1

We selected from the output of the IEDB MHC-I processing tool for each protein only those peptides with a predicted IC50$\leq$ 500 nM and plotted the Total Scores of the peptides from each protein for each allele. The controlling alleles were predicted to bind a large number of peptides with a range of total scores. The majority of the peptides came from Env and Pol and the lowest number of peptides from Rev and Tat, with the exception of HLA-B*27:05.

The non-controlling alleles HLA-B*35:03 and -B*55:01 were predicted to present

very few peptides in total, with much lower total scores than the controlling alleles. The non-controlling alleles HLA-B*07:02 and -B*18:01 however were predicted to bind peptides with a similar range of total scores to the controlling alleles, with a similar total number being presented.

One conclusion that can be reached when using this method is that some alleles are associated with fast progression because they simply do not present enough epitopes to result in a successful immune response. This explanation can be applied to HLA-B*35:03 and -B*55:01 but not HLA-B*07:02 or -B*18:01. However, the issue with the method of using a 500 nM IC50 threshold cut-off is that we cannot necessarily compare predicted IC50 values between alleles[142][143], as the percentage of peptides predicted to bind with 500 nM or lower varies between alleles. Therefore, a fairer comparison would be to take the top 1% of binding peptides, so that each allele binds the same number of peptides.

5.2.3 Results: method 2

We selected from the output of the IEDB MHC-I processing tool those peptides with a Total Scores within the top 1% of all predicted peptides for the HIV-1 proteaome. Using the top 1% cut-off instead of the 500 nM affinity cut-off means that the number of peptides bound by each allele is the same, and results in the number of peptides predicted to bind to the controlling alleles is actually reduced and the non-controlling alleles HLA-B*53:01 and -B*35:03 are predicted to bind as many peptides as the non-controlling alleles. Therefore, we can no longer claim that these alleles are associated with long-term non-progression because they do not bind enough peptides.

However, when comparing progressors to LTNPs, we found that the average Total Score for the peptides presented by the non-controlling alleles HLA-B*35:03 and -B*55:01 was much lower for each protein than the controlling alleles (Figure 5.4 A&B), which suggests that these alleles present fewer immunogenic epitopes than the controllers, and so are less able to control the spread of the virus. However, the non-controlling HLA-B*07:02 allele was predicted to bind a similar number of Pol, Env and Gag peptides. Furthermore, HLA-B*18:01 was predicted to bind a similar number of peptides overall as the controlling alleles, and with a similar range of Total Scores.

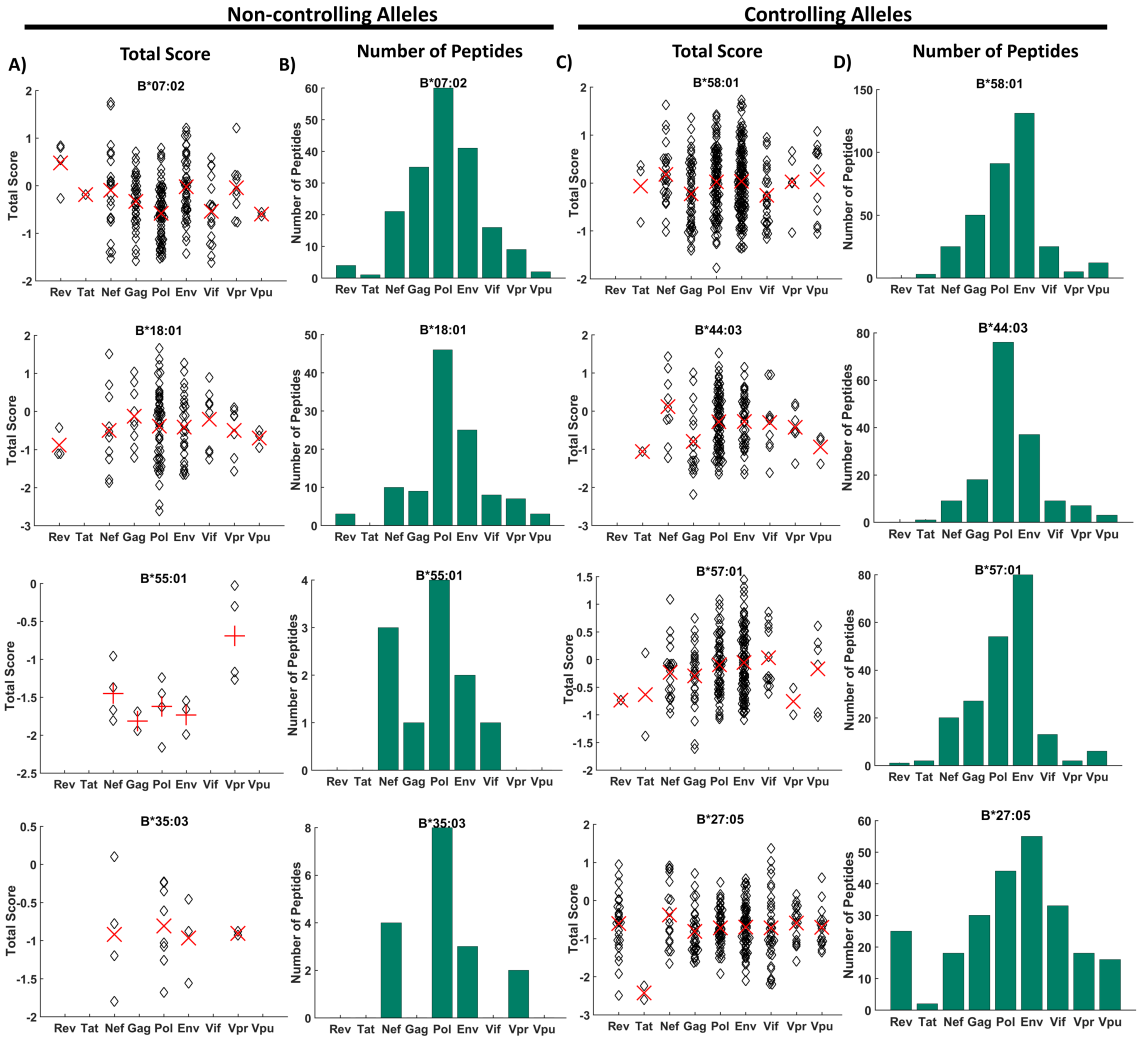


Figure 5.3: IEDB Total Scores for Threshold Method The IEDB MHC-I processing tools were used to analyse the distribution of only those HIV-1-derived peptides with predicted IC50<500 nM for A) non-controlling alleles B*07:02, B*18:01, B*55:01 and B*35:03 and C) controlling alleles, HLA-B*58:01, B*44:03, B*57:01 and B*27:05. The IEDB Total Score for each peptide is plotted according to which protein they originate from. The red crosses indicate the average total score of peptides from each protein. B,D) The number of peptides with IC50<500 nM is compared for each protein.

Again, we would have expected Gag peptides to have the highest average Total Scores, however, the average Gag Total Score was one of the lowest for the controlling alleles, with the highest average scores coming from Pol, Env, Nef and Vif in general (Figure 5.4 C). Similarly, for the controlling alleles HLA-B*58:01, -B*44:03 and -B*57:01, Pol and Env were the two proteins predicted to produce the largest number of binding peptides to the controlling alleles (Figure 5.4 D), and not Gag.

Focussing on Gag-derived peptides alone, we found that the highest average Total Score is associated with the controlling allele HLA-B*58:01, and the lowest average Total Score is associated with the non-controlling allele HLA-B*35:03. However, again we found no obvious distinction between the average Gag peptide Total Scores between the chosen set of controlling and non-controlling alleles that could explain their observed differences in rates of disease progression. In fact, from these predictions, we would expect Pol peptides to control HIV progression, as Pol is a highly conserved sequence and yields a large number of peptides with high Total Scores. In contrast, the Env sequence is highly variable[144], so even though it also produces many peptides with high Total Scores, the higher probability of escape mutations reduces its immunogenicity.

Borghans et al.[141] used NetMHC3.0, a peptide-MHC binding prediction tool, available at `www.cbs.dtu.dk/services/NetMHC-3.0/` to predict the affinities of HIV peptide to different HLA alleles, and then ranked each peptide among all HIV peptides from highest to lowest affinity, where a lower rank corresponds to a high affinity binder. They compared the predicted ranks of peptides from different HIV-1 proteins between a group of controlling alleles (HLA-B*27:05, -B*57:01 and -B*58:01) and a group of non-controlling alleles (HLA-B*35:03 and -B*53:01). They then looked at only the top 3 ranking Gag epitopes and found the controlling group had a significantly higher preference for Gag than the non-controlling group. Doing the same for the other proteins they found the non-controlling group had a preference for Nef peptides over the controlling group. Significant differences in the preferences for Vpu, p17, Vif and Ref epitopes were also observed but it was concluded that the median ranks of these peptides were so high (i.e. the peptide affinities were so low) that the differences were most likely not physiologically important.

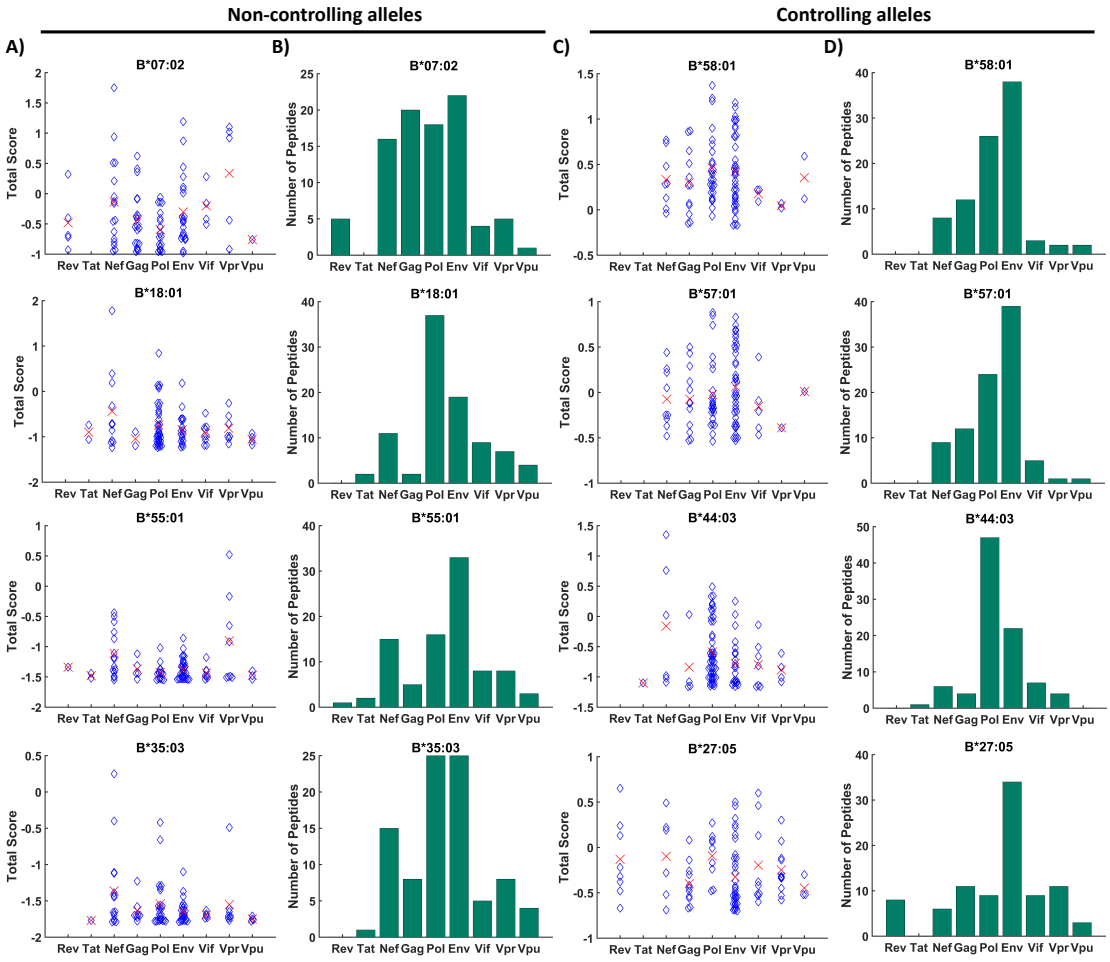


Figure 5.4: The IEDB prediction tool suggests that Pol and Env produce the majority of peptides presented on HLA molecules. The IEDB MHCI processing tools were used to analyse the distribution of the top 1% of HIV-1-derived peptides predicted to be presented on MHCI molecules. The predictions were made for controlling alleles, HLA-B*58:01, B*44:03, B*57:01 and B*27:05, and non-controlling alleles B*07:02, B*18:01, B*55:01 and B*35:03. A,C) The IEDB Total Score for each peptide is plotted according to which protein they originate from. The red crosses indicate the average total score of peptides from each protein. B,D) The number of peptides in the top 1% is compared for each protein.

We carried out a similar analysis to that in Borghans et al. [141] on the top 1% of peptides. We compared the median predicted Total Score of all peptides in the top 1% of each allele, grouping by control vs non-control, using a Wilcoxon rank sum test (equivalent to the Mann-Whitney U-test used by Borghans et al.[141]). The analysis revealed a statistically significant difference for all HIV proteins considered, suggesting that controlling alleles prefer HIV peptides from Nef ($p = 3.6 \times 10^{-6}$), Gag ($p = 2.4 \times 10^{-6}$), Pol ($p = 3.1 \times 10^{-18}$), Env ($p = 1.5 \times 10^{-24}$), Vif ($p = 1.2 \times 10^{-5}$, Vpr ($p =$

0.0058) and Vpu ($p = 1.0 \times 10^{-4}$) in general compared to non-controlling alleles (Rev and Tat were not included due to the low numbers of peptides from these proteins in the top 1%). We could conclude from this analysis that controlling alleles prefer to present peptides from the entire HIV genome in general and that we would expect the two proteins with the lowest p-values, Pol and Env to be associated with control of HIV, which does not provide any explanation for the immunodominance of Gag epitopes.

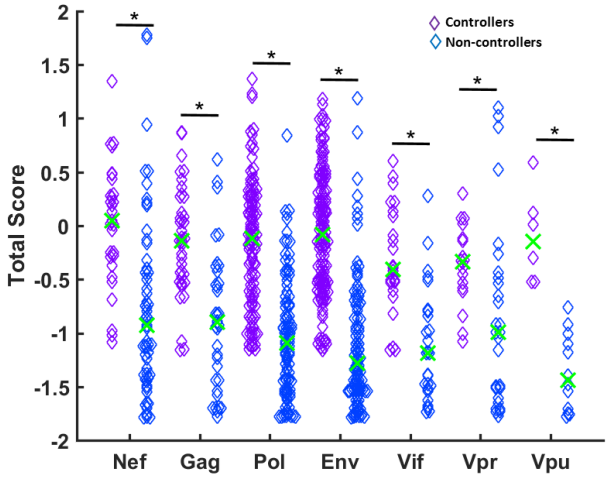


Figure 5.5: Total Scores of top 1% of HIV epitopes from different HIV proteins grouped by controllers vs non-controllers The median Total Score of the top 1% predicted HIV peptides for either controlling alleles or non-controlling alleles was compared using a Wilcoxon rank sum test. A significantly higher median Total Score when binding to the controlling group was observed for all HIV proteins included in the analysis, suggesting controlling alleles preferentially bind HIV peptides in general compared to non-controlling alleles.

5.3 A Mechanistic Model of HIV-1 Antigen Presentation

When a cell is infected by a HIV-1 virion, its constituent proteins enter the cytoplasm, and become available for degradation by the proteasome. Following translocation to the endoplasmic reticulum (ER) via the transported associated with antigen processing (TAP) these peptides are then available for loading on to MHC class I molecules and presentation at the cell surface.

During viral infection, a cytokine known as interferon gamma (IFN-γ) is produced, which up-regulates the cell surface expression of MHC, suggesting that the cell surface abundance of a peptide-MHC complex is an important factor in determining the T-cell

response[6] (discussed in more detail in Chapters 1 & 3). The abundance of a peptide-MHC complex on the cell surface is dependent upon the peptide-MHC affinity, but also the abundance of the peptide in the ER. The abundance of the peptide in the ER is dependent upon the proteasomal cleavage and TAP transport, as accounted for by the MHC-I processing tool, however the abundance of the protein from which the peptide is cleaved is also very important. The IEDB MHC-I processing tool cannot account for this abundance because it is just dealing with the protein sequences. Therefore, we sought to put together a model that predicts the cell surface abundance of HIV peptides by different HLA alleles by combining the MHC-I processing tool predictions with a model for the HIV protein abundances in the cytoplasm.

There are two sources of HIV-1 peptides; HIV-1 virion proteins and HIV-1 proteins synthesised by the host cell machinery during viral replication. Presentation of virion derived peptides provides the immune system a chance to detect and destroy infected cells before viral replication begins in earnest. Once the HIV genome has been integrated in to the host genome, the host cellular machinery will synthesise HIV-1 proteins in the cytoplasm which begins the replication process. These proteins will also be degraded in to peptides and some of these peptides will be presented on the cell surface. The HIV-1 host-synthesised proteins will accumulate to much higher levels in the cytoplasm than the virion proteins, and these high concentrations will be sustained during replication. Presentation of these peptides will most likely be of higher abundance than the virion peptides, and their presentation will be sustained for as long as replication occurs. This gives the immune system a second opportunity to recognise and destroy an infected cell.

We therefore sought to construct a model that incorporates the effects of viral protein intracellular kinetics, protein sequence specificities and MHC-I binding to produce dynamic predictions of the antigen presentation profile of infected cell. We can model the dynamics of HIV-1 peptides deriving from host-synthesised viral proteins deterministically using a set of differential equations to describe the change in copy number of each species involved in the process. However, due to the low copy number of HIV-1 proteins within a virion, we can not assume spatial homogeneity of molecular concentrations or constant rates of reaction. Therefore, we must model virion peptide presentation stochas-

tically, and as a result we constructed two models of HIV-1 antigen presentation; a virion model and a HIV-1 replication model as shown diagrammatically in Figure 5.6.

5.3.1 Methods: constructing the HIV-1 replication model

To create an integrated model of viral infection, *de novo* viral protein synthesis and peptide presentation on MHC class I molecules, we started by combining three existing models of HIV-1 intracellular kinetics, Kim & Yin[117], Reddy & Yin[136] and Wang & LuHua[145] to create a more complete model. We decided to use existing models and combine them, instead of putting together a model of our own, as the existing models have been peer reviewed, use experimentally derived values for important parameters, and are detailed enough on their own that once combined we are able to model the kinetics of almost the entire HIV proteaome. Hwijin Kim & John Yin presented two papers

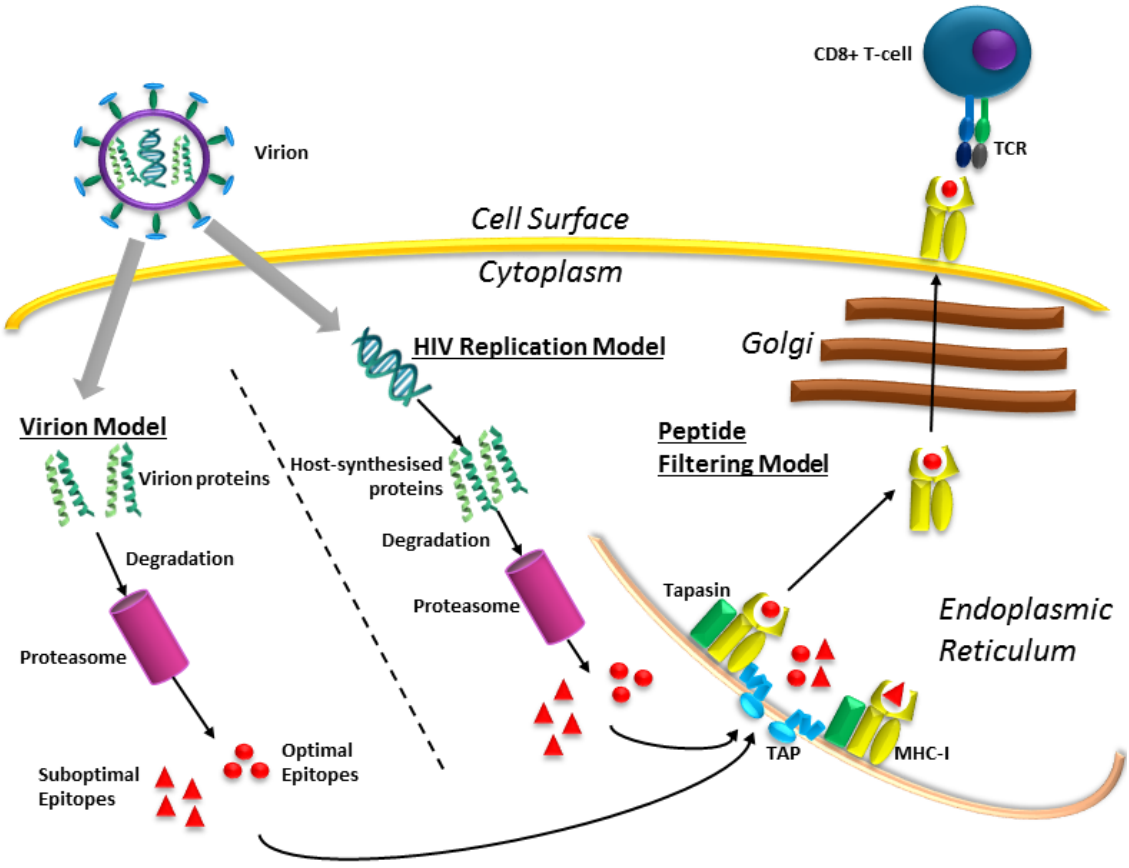


Figure 5.6: Combined model of HIV-1 infection and cell surface peptide presentation on MHCI molecules. Diagrammatic representation of the combination of the separate models that comprise the combined model: the HIV kinetics models[117, 136, 145], and the peptide filtering model[22].

in 2005[117, 146], one which simulated HIV-1 kinetics from transcription in the nucleus to study the effects of Rev on RNA splicing, whilst the other studied the purpose of the feedback loops of Rev and Tat on the growth of HIV-1. Reddy & Yin's 1999[147] model covers the process from reverse transcription, which occurs straight after viral entry in to the host cell, all the way to the budding and maturation of new virions. Wang & LuHua[145] modelled the kinetics of Vif and Gag synthesis and well as the budding process. This model is the most recent and up to date of the three and so its budding kinetics were used instead of Reddy & Yin's in the combined model. Combining all three models we were able to account for the intracellular kinetics of Rev, Tat, Gag, GagPol, Env and Vif. The model was extended by inserting equations for the remaining proteins Nef, Vpr and Vpu. We found half-life data from the literature from which we calculated their degradation rates. We estimated their translation probabilities so that they produce sufficient levels of proteins to allow formation of virions with the correct levels of proteins in them, as measured experimentally.

In order to predict the cell surface presentation, these models must be combined with the peptide filtering model[22], however we chose to simulate the virion model stochastically and the intracellular kinetics model deterministically (see Figure 5.6). We simulated the virion model stochastically due to the low copy number of proteins within a HIV virion, which will mean that the rate of reactions are not constant as the distribution of molecules of each species cannot be assumed to be homogeneous, and therefore a deterministic approach would be inaccurate.

5.3.2 Modelling HIV-1 intracellular kinetics

The combined model of HIV-1 intracellular kinetics begins with the Kim & Yin model[117]. The differential equations describing the kinetics of each species in time are as follows.

To begin with consider the kinetics of the full-length mRNA in the nucleus F_N.

$$\frac{d[F_N]}{dt} = T_{C_b} + T_{C_{add}} \frac{K_{Tat}[T_N]}{1 + K_{Tat}[T_N]} pv + k_d^{(1)}[FR_N^{(1)}] - (k_{sp}^F + k_{deg,N}^{RNA} + k_a^1[R_N])[F_N] \quad (5.1)$$

The full-length HIV-1 mRNA is produced first via transcription from the HIV-1 genome

at the basal rate of host transcription T_{C_b}. Once the protein Tat, T_N, has been translated and imported back in to the nucleus, it binds with the Tat responsive element (TAR), which along with some cellular co-factors, dramatically increases the transcription rate of all mRNAs. This increase in transcription via Tat binding is denoted $T_{C_{add}}$. Michaelis-Menten kinetics are used to account for the saturation of either the number of polymerases or in the number of the host cellular co-factors. In the Michaelis-Menten expression, K_{Tat} is the equilibrium constant of Tat with TAR. The increase in transcription due to Tat binding is multiplied by factor pv, which is the number of proviruses that have been integrated in to the genome.

The term $k_d^{(1)}[FR_N^{(1)}]$ describes the dissociation of a Rev protein from the complex $FR_N^{(1)}$, where FR_N denotes the nuclear full-length transcript with Rev bound, and the superscript of (1) denotes that one Rev protein is bound to the complex. Therefore the dissociation of a single Rev from $FR_N^{(1)}$ creates a F_N. The splicing of full-length nuclear transcript in to singly-spliced mRNA is denoted by the rate coefficient k_{sp}^F, and the degradation of the full-length mRNA in the nucleus is denoted $k_{deg,N}^{RNA}$. The term $k_a^{(1)}[R_N][F_N]$ represents the binding of a single Rev R_N to a free full-length nuclear transcript, where $k_a^{(1)}$ is the rate coefficient for the association rate of one Rev to the full-length mRNA, thus producing $FR_N^{(1)}$.

Rev-bound full-length mRNA, $FR_N^{(i)}$, where i is the number of Rev proteins bound to the transcript, behaves slightly differently from Rev-free full-length mRNA:

$$\begin{aligned}\frac{d[FR_N^{(i)}]}{dt} &= k_a^{(i)}[R_N][FR_N^{(i-1)}] + k_d^{(i+1)}[FR_N^{(i+1)}] - (k_d^{(i)} \\ &+ k_a^{(i+1)}[R_N] + k_{exp}^{(F,i)} + (1-d^{F,(i)})k_{sp}^F + k_{deg,N}^{RNA})[FR_N^{(i)}]\end{aligned} \tag{5.2}$$

for $i = 1, 2, ..., sn$.

The first term in Equation 5.2 describes the creation of $FR_N^{(i)}$, via the association of a single Rev protein R_N to a full-length transcript with $i-1$ Rev proteins bound, with rate constant $k_a^{(i)}$. Similarly, $FR_N^{(i)}$ can be created via the dissociation of a single Rev protein from $FR_N^{(i+1)}$, with rate constant $k_d^{(i+1)}$. On the other hand, a Rev protein can unbind to $FR_N^{(i)}$ with dissociation rate $k_d^{(i)}$, which creates $FR_N^{(i-1)}$. Similarly, if single Rev protein

binds to $FR_N^{(i)}$, with association rate $k_a^{(i+1)}$ this creates the complex $FR_N^{(i+1)}$. $FR_N^{(i)}$ can also be spliced with rate $(1-d^{S,(i)})k_{sp}^F$, where $d^{S,(i)}$ is the delay in splicing due to the presence of the bound Rev proteins. $FR_N^{(i)}$ is exported to the cytoplasm with rate $k_{exp}^{F,(i)}$, and is degraded with rate constant $k_{deg,N}^{RNA}$.

Now we will look at the kinetics of the singly-spliced mRNA in the nucleus:

$$\frac{d[S_N]}{dt} = k_{sp}^F[F_N] + k_d^{(1)}[SR_N^{(1)}] - (k_{sp}^S + k_{deg,N}^{RNA} + k_a^{(1)}[R_N])[S_N] \tag{5.3}$$

Full-length mRNA is spliced in the nucleus with rate constant k_{sp} to produce singly-spliced mRNA, S_N. A single Rev protein can dissociated from a $SR_N^{(1)}$ complex, which is singly-spliced nuclear mRNA with a single Rev protein attached, at rate k_d, which creates a free S_N. Singly-spliced mRNA can also be spliced with rate constant k_{sp}^S in to multiply-spliced mRNA, and S_N is also degraded with rate $k_{deg,N}^{RNA}$. The concentration of S_N is reduced when a nuclear Rev protein R_N associates with S_N with rate coefficient $k_a^{(1)}$, thus producing $SR_N^{(1)}$.

Similar to $FR_N^{(i)}$, the kinetics of singly-spliced mRNA with i bound Rev proteins, $SR_N^{(i)}$ are different to those of S_N,

$$\begin{aligned}\frac{d[SR_N^{(i)}]}{dt} = k_a^{(i)}[R_N][SR_N^{(i-1)}] + k_d^{(i+1)}[SR_N^{(i+1)}] + (1-d^{F,(i)})k_{sp}^F[FR_N^{(i)}] - (k_d^{(i)} \\ +k_a^{(i+1)}[RN] + k_{exp}^{S,(i)} + (1-d^{S,(i)})k_{sp}^S + k_{deg,N}^{RNA})[SR_n^{(i)}]\end{aligned} \tag{5.4}$$

for $i = 1,2,...,sn$.

$SR_N^{(i)}$ is created when a Rev protein R_N associates with a singly-spliced mRNA molecule with $i-1$ Revs bound, with rate coefficient $k_a^{(i)}$. Similarly, $SR_N^{(i)}$ is also created if a single Rev protein dissociates with rate constant $k_d^{(i+1)}$ from singly-spliced mRNA with $i+1$ bound Rev molecules. $SR_N^{(i)}$ can also be created following the splicing of $FR_N^{(i)}$ with rate coefficient $(1-d^{F,(i)})k_{sp}^F$. A Rev protein can dissociate from $SR_N^{(i)}$ with rate constant $k_d^{(i)}$, which creates $SR_N^{(i-1)}$, and a Rev protein can also associate with $SR_N^{(i)}$ with rate constant $k_a^{(i+1)}$ to produce $SR_N^{(i+1)}$. Once a threshold number of Rev proteins are bound, $SR_N^{(i)}$ can be exported to the cytoplasm with rate constant $k_{exp}^{S,(i)}$. $SR_N^{(i)}$ can also be spliced with rate constant $(1-d^{S,(i)})$ to create multiply spliced mRNA, M_N and i free Rev proteins.

The differential equation for the kinetics of multiply-spliced mRNA in the nucleus, M_N is written as:

$$\frac{d[M_N]}{dt} = k_{sp}^{S}[S_N] + \sum_{i=1}^{sn}((1-d^{S,(i)})k_{sp}^{S}[SR_N^{(i)}]) - (k_{exp}^{M} + k_{deg,N}^{RNA})[M_N] \tag{5.5}$$

Multiply spliced mRNA is produced via the splicing of Rev-free singly-spliced mRNA S_N and also via the splicing of singly-spliced mRNA with i Rev proteins bound, $SR_N^{(i)}$, and the rate constant for both is k_{sp}^{S}. The maximum number of Rev proteins that can bind to a single transcript is denoted sn. The second term on the right-hand side also includes the factor $(1-d^{S,(i)})$, where $d^{S,(i)}$ is the delay in splicing due to the presence of the bound Rev proteins. As with the previous two equations, M_N degrades with rate constant $k_{deg,N}^{RNA}$. The term $k_{exp}^{M}[M_N]$ accounts for the fact that multiply-spliced mRNA can be exported to the nucleus independent of Rev, with rate constant k_{exp}.

The kinetics of cytoplasmic multiply-spliced mRNA, M_C, are as follows,

$$\frac{d[M_C]}{dt} = k_{exp}^{M}[M_N] - k_{deg,C}^{RNA}[M_C] \tag{5.6}$$

where M_C is degraded with rate constant $k_{deg,C}^{RNA}$ in the cytoplasm.

As previously mentioned, once a threshold number of Rev molecules have associated with either full-length or singly-spliced mRNA, they too can be exported to the cytoplasm, and the differential equation describing their cytoplasmic kinetics are as follows:

$$\frac{d[F_C]}{dt} = \sum_{i=1}^{sn}(k_{exp}^{F,(i)}[FR_N^{(i)}]) - k_{deg,C}^{RNA}[F_C] \tag{5.7}$$

$$\frac{d[S_C]}{dt} = \sum_{i=1}^{sn}(k_{exp}^{S,(i)}[SR_N^{(i)}]) - k_{deg,C}^{RNA}[S_C] \tag{5.8}$$

$SR_N^{(i)}$ and $FR_N^{(i)}$ can only be exported to the cytoplasm once a threshold level of Rev proteins has bound to the transcript. Therefore for $i < threshold$, $k_{exp}^{X,(i)} = 0$ but for $i \geqslant threshold$, $k_{exp}^{X,(i)} \neq 0$. The threshold number of Rev proteins that are required to bind before nuclear export can begin is given in [117] as 7, with the maximum number of Rev proteins that can bind to a single transcript, sn, being 12.

Once they enter the cytoplasm, the i bound Rev proteins instantaneously unbind. The kinetics of Rev proteins in the cytoplasm, R_C are as follows,

$$\frac{d[R_C]}{dt} = f_{Rev} \cdot Tr \cdot f_{rev}^{M}[M_C] + k_{exp}^{R}[R_N] + \sum_{i=1}^{sn}(i \cdot (k_{exp}^{F,(i)}[FR_N^{(i)}] + k_{exp}^{S,(i)}[SR_N^{(i)}])) - (k_{imp}^{R} + k_{deg,C}^{R})[R_C] \tag{5.9}$$

Rev is produced in the cytoplasm via translation, with rate constant Tr, from the fraction of multiply-spliced mRNA, f_{rev}^{M}, that encodes for the rev mRNA, where f_{Rev} is the probability for this rev mRNA to encode for the Rev protein. For Rev to be able to enable the export of full-length and singly-spliced mRNA, it has to be imported in to the nucleus, which occurs with rate coefficient k_{imp}^{R}. Nuclear Rev R_N can also be exported back to the cytoplasm independently of any mRNA transcript, with rate constant k_{exp}^{R}. Those mRNA nuclear transcripts with a threshold level of Rev proteins bound are transported in to the cytoplasm with rate constant $k_{exp}^{X,(i)}$ where X stands for either full-length, F, or singly-spliced, S, and they then release their bound Rev proteins back in to the cytoplasm. Therefore, for an exported concentration of transcript X with i bound Rev proteins, where $i \geqslant threshold$, there will be $(i \cdot [XR_N^{(i)}])$ Rev proteins released. To account for the contribution to the cytoplasmic Rev pool from all exported transcripts with $i \geqslant threshold$ bound Rev proteins we must sum up over i, where $k_{exp}^{X,(i)}$ for $i < threshold$ is zero. The degradation of Rev in the cytoplasm occurs with rate constant $k_{deg,C}^{R}$.

Once transported in to the nucleus, the kinetics of the Rev protein, R_N, are as follows:

$$\frac{d[R_N]}{dt} = k_{imp}^{R}[R_C] + \sum_{i=1}^{sn}(k_d^{(i)} \cdot ([FR_N^{(i)}] + [SR_N^{(i)}])) + \sum_{i=1}^{sn}(i \cdot k_{deg,N}^{RNA}([FR_N^{(i)}] + [SR_N^{(i)}])) + \sum_{i=1}^{sn}((1 - d^{S,(i)})k_{sp}^{S,(i)}[SR_N^{(i)}]) - \Big(\sum_{i=1}^{sn}(k_a^{(i)}([FR_N^{(i)}] + [SR_N^{(i)}]) + k_{exp}^{R} + k_{deg,N}^{R}\Big)[R_N] \tag{5.10}$$

Rev can be imported in to the nucleus with rate constant k_{imp}^{R} and can be exported back in to the cytoplasm with rate constant k_{exp}^{R}. When a Rev dissociates with rate constant $k_d^{(i)}$ from a full-length mRNA $FR_N^{(i)}$, or a singly-spliced mRNA $SR_N^{(i)}$, with i bound Rev proteins, it increases the nuclear pool of Rev. Therefore, we must sum over all i from 1

to the maximum number of bound Rev, *sn*, to account for the dissociation of Rev proteins from all Rev-bound transcripts. When a Rev-bound transcript, $XR_N^{(i)}$ degrades with rate constant $k_{deg,N}^{RNA}$, it releases the i Rev proteins it is bound to, and so we must also sum up over all i to account for this. Similarly, we must sum over all i when considering the i Rev proteins released during the splicing of singly-spliced mRNA, at rate constant $(1 - d^{S,(i)})k_{sp}^{S,(i)}$. We must also sum over all association events, where a Rev protein associates with a transcript $XR_N^{(i-1)}$, with rate constant $k_a^{(i)}$. Finally, Rev degrades in the nucleus with rate constant $k_{deg,N}^{R}$.

The cytoplasmic kinetics of the protein Tat which is required for the increase in the transcription rate of HIV transcripts, are described in Equation 5.11. Tat is translated from both multiply-spliced mRNA, M_C and singly-spliced mRNA, S_C, at rate Tr where the fraction of tat mRNA in these transcripts is f_{tat}^{M} and f_{tat}^{C} respectively. The probability that this tat mRNA encodes for Tat is given by f_{Tat}.

$$\frac{d[T_C]}{dt} = f_{Tat} \cdot Tr\dot{(}f_{tat}^{S}[S_C] + f_{tat}^{M}[M_C]) + k_{exp}^{T}[T_N] - (k_{imp}^{T} + k_{deg,C}^{T})[T_C] \tag{5.11}$$

Cytoplasmic Tat is imported in to the nucleus with rate constant k_{imp}^{T} where it can increase the rate of transcription. Free Tat proteins in the nucleus, T_N can also be exported back in to the cytoplasm with rate constant k_{exp}^{T}. Tat is degraded in the cytoplasm with rate constant $k_{deg,C}^{T}$.

Once in the nucleus, the kinetics of Tat are as follows,

$$\frac{d[T_N]}{dt} = k_{imp}^{T}[T_C] - (k_{exp}^{T} + k_{deg,N}^{T})[T_N] \tag{5.12}$$

where k_{exp}^{T} is the rate constant of Tat export in to the cytoplasm, and $k_{deg,N}^{T}$ is the rate constant for degradation of nuclear Tat.

The Kim and Yin[117] model as described above provides the kinetics for the steps of HIV-1 transcript synthesis, and the translation and kinetics of the proteins Rev and Tat. To model the dynamics of the important structural proteins, Gag, GagPol and Env, we use the model presented by Reddy and Yin[147]. They write the differential equations

describing the protein dynamics as follows:

$$\frac{d[Gag]}{dt} = f_{Gag} \cdot Tr[F_C] - k_{Gag}[Gag] - k_{bud}[Gag] \tag{5.13}$$

$$\frac{d[GagPol]}{dt} = f_{GagPol} \cdot Tr[F_C] - k_{GagPol}[GagPol] - k_{bud}[GagPol] \tag{5.14}$$

$$\frac{d[Env]}{dt} = f_{Env} \cdot Tr[S_C] - k_{Env}[Env] - k_{bud}[Env] \tag{5.15}$$

Both Gag and GagPol are translated from full-length cytoplasmic mRNA, F_C, with translation probability f_{Gag} and f_{GagPol} respectively. Env is translated from singly-spliced cytoplasmic mRNA, S_C, with translation probability f_{Env}. The degradation of each protein is described by the second term in each of the Equations 5.13-5.15, where k_{Prot} is the degradation rate of each protein Gag, GagPol and Env. The third term in Equations 5.13-5.15 describes the budding process, where the proteins are translocated to the cell membrane to create new virions. This step is not included in the Reddy and Yin model, but is taken from Wang and LuHuas[145] model where the value of k_{bud} was chosen to ensure the levels of full-length cytoplasmic mRNA reached a steady state at $3,900$ molecules per cell. This steady state level is the average of a set of experimental data measuring the concentration of intracellular Gag.

The Wang & Hua[145] model also included the synthesis of the Vif protein, whose dynamics are as follows:

$$\frac{d[Vif]}{dt} = f_{Vif} \cdot Tr[S_C] - k_{Vif}[Vif] - k_{bud}[Vif] \tag{5.16}$$

Vif is translated from single-spliced mRNA with probability f_{Vif}, degrades with rate constant k_{Vif} and is translocated to the cell membrane for budding with k_{bud}.

Unfortunately, none of the previously existing models include the kinetics of the remaining proteins Nef, Vpr, Vpu. Nef is translated from multiply-spliced cytoplasmic mRNA and is incorporated in to a single virion at between $60-200$ copies per virion. To model Nef we used the average rate of degradation of eukaryotic proteins as given in

[136] for the degradation rate, and set the budding rate to be the same as the rate given in [145] for the previous equations. We then set the probability of translation f_{Nef} so that the trade-off between it and the two other rates k_{Nef} and k_{bud} resulted in there being between $60-200$ copies of Nef in a single budding virion[148]. The Vpr protein is relatively stable, with a measured half-life of around $20h$[149], which we used to calculate the degradation rate of Vpr, k_{Vpr}. Vpr is found in the HIV-1 virion at a ratio of $1:7$ to the number of copies of Gag[135], and so we set the probability of translation f_{Vpr} from singly-spliced mRNA to ensure that the ratio of Gag:Vpr is as stated, using the value calculated from the half-life for the degradation rate, and the same budding rate as the other proteins. Finally, for the Vpu protein, we could not located any measurements of Vpu half-life in the literature, nor could we find any information regarding the number of copies of Vpu in a HIV-1 virion. The mRNA and protein dynamics from the combined

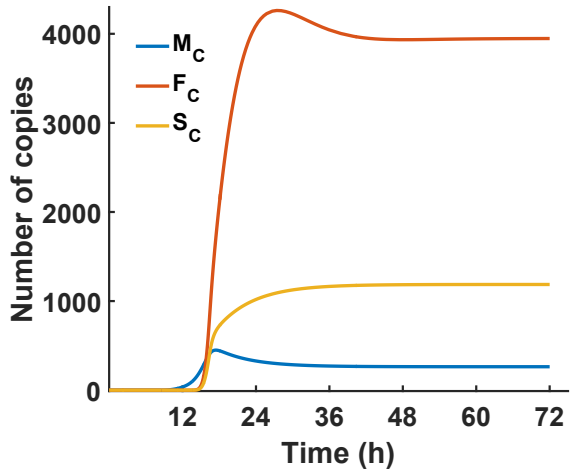


Figure 5.7: Simulated HIV mRNA. The combined model was used to simulate full-length (F_C), singly-spliced (S_C) and multiply-spliced (M_C) mRNA copies in the cytoplasm. The full-length cytoplasmic mRNA (F_C) reaches a steady state level of $3,900$ copies, which agrees with the experimentally measured average[150].

and more complete model of HIV intracellular kinetics are shown in Figures 5.7 & 5.9 A respectively. Following viral entry to a cell, there will be a delay before the viral genome is integrated in the cell genome in the nucleus, experimentally measured to be around 8.5 hours post infection[151], during which time the model will only simulate the degradation of the virion proteins and their peptide production. Gag is the most abundant protein for the majority of the replication process, with the exception of the short time required for the translation of Rev from multiply-spliced mRNA and its subsequent binding to full-

length and singly-spliced mRNA in the nucleus. During this time, the concentration of Rev and Tat available for cytoplasmic degradation will be very low due to the shuttling back and forth from the cytoplasm to the nucleus, and the binding to the other HIV transcripts. The Nef protein however, which is also transcribed from multiply-spliced mRNA accumulates in the cytoplasm, as indicated by the yellow curve in Figure 5.9 A. Gag begins to accumulate to very high numbers at around 15 hours post infection. Pol, Env and Vpr also accumulate to considerable numbers during this time.

Table 5.1: HIV-1 intracellular kinetics model parameters.

Parameter	Value	References
Basal transcription rate, T_{C_b}	4.1667×10^{-3} transcripts/s	[117, 136]
Increase in Transcription by Tat Transactivation, $T_{C_{add}}$	0.4125 transcripts/s	[152, 153, 154]
Equilibrium dissociation constant of Tat with TAR, K_{Tat}	5.2453×10^{-5}/molecule	$28.57/\mu M$[145]
Splicing rate constants, $k_{sp}^F = k_{sp}^S$	6.95×10^{-4}molecules^{-1}s^{-1}	[155, 117, 136]
Rate of export from the nucleus $k_{exp}^{F,(i)} = k_{exp}^{S,(i)} = k_{exp}^M = k_{exp}^R, k_{exp}^T = 0$	5.7833×10^{-4}s^{-1}	[156]
Rate of Translation k_{Trans}	$0.075s^{-1}$	[157, 117, 136]
Fraction of Full-Length mRNA that encodes for Gag f_{Gag}	0.95	[136]
Fraction of Full-Length mRNA that encodes for Gag-Pol f_{GagPol}	0.05	[136]
Fraction of Singly-spliced mRNA that encodes for Env f_{Env}	0.15	[136]
Fraction of Multiply-spliced mRNA that encodes for Rev f_{Rev}^M	0.19	[117]

Continued on next page

Table 5.1 – continued from previous page

Parameter	Value	References
Probability that Rev mRNA will encode for Rev f_{Rev}	0.5	[117]
Fraction of Multiply-spliced mRNA that encodes for Tat mRNA f_{Tat}^{M}	0.05	[117]
Fraction of Singly-spliced mRNA that encodes for Tat mRNA f_{Tat}^{S}	0.05	[117]
Probability that Tat mRNA will encode for Tat f_{Tat}	1.0	[117]
Splicing Delay Factor Due to Rev $d^{F,(i)} = d^{S,(i)}$	0.8	[117]
Nuclear Import Rate Constant $k_{imp}^{T} = k_{imp}^{R}$	$5.7833 \times 10^{-3} s^{-1}$	[117, 158]
Degradation rate for Rev in the cytoplasm, $k_{deg,C}^{Rev}$	$4.833 \times 10^{-6} s^{-1}$	[117]
Degradation rate for Rev in the nucleus $k_{deg,N}^{Rev} = k_{deg,N}^{Prot}$	$1.2 \times 10^{5}\mathrm{s}^{1}$	[117]
Degradation rate for Tat in the cytoplasm and nucleus $k_{deg,C}^{Tat} = k_{deg,N}^{Tat}$	$4.278 \times 10^{5}\mathrm{s}^{-1}$	[117]
Degradation rate for gp120 (Env Precursor) in cytoplasm $k_{deg,C}^{gp120}$	$5.55 \times 10^{-6} s^{-1}$	[136]
Association constant for Rev with RRE, $k_a^{(1)}$ and $k_a^{(i)}$	$k_a^{(1)} = 0.0132\mathrm{molecules}^{1}\mathrm{s}^{1}$, $k_a^{(i)} = 0.0233\ \mathrm{molecules}^{1}\mathrm{s}^{1}$	[117]
Dissociation constant for Rev with RRE, $k_d^{(1)}$ and $k_d^{(i)}$	$k_d^{(1)} = 3.0 \times 10^{5}\ \mathrm{molecules}^{1}\mathrm{s}^{1}$, $k_d^{(i)} = 3.8 \times 10^{2}\mathrm{molecules}^{1}\mathrm{s}^{1}$	[117]

Continued on next page

Table 5.1 – continued from previous page

Parameter	Value	References
Rate of export of proteins through budding k_{bud}	0.08 h^{-1}	[145]
Protein degradation of eukaryotic proteins k_{deg}^{prot}	$1.083 \times 10^4 s^1$	[136]
Fraction of Singly-spliced mRNA that encodes for Vif f_{Vif}	0.1	NA
Fraction of Singly-spliced mRNA that encodes for Vpr f_{Vpr}	0.27	NA
Fraction of Singly-spliced mRNA that encodes for Vpu f_{Vpu}	0.1	NA
Fraction of Multiply-spliced mRNA that encodes for Nef f_{Nef}	0.5	NA
Degradation rate of Gag, k_{Gag}	0.1054 h^{-1}	[136]
Degradation rate of GagPol, k_{Pol}	0.1159 h^{-1}	NA
Degradation rate of Vpr, k_{Vpr}	0.0346 h^{-1}	[149]
Degradation rate of Env, k_{Env}	0.02 h^{-1}	[136]
Degradation rate Vpu, k_{Vpu}	0.086 h^{-1}	NA

5.3.3 Estimating the peptide-MHC-I unbinding rates

The IEDB MHC-I binding predictor tool[76] predicts the affinity of the binding between a peptide sequence and a chosen MHC-I allele. We used consensus HIV clade C proteome `http://www.hiv.lanl.gov` as inputs to the IEDB predictor. The output of the tool is a large table of peptide sequences derived from the input protein sequence, and the predicted affinity between the peptide and the HLA allele.

The affinity is given as an IC50 (nM) value, or half the maximal inhibitory concentra-

tion. This value is found by determining the concentration of test peptide required to fill half of the MHC binding sites when competing against a labelled peptide. The machine learning algorithm is then built upon a data set of these values and their corresponding peptide sequences.

In Section 5.2.1 we used two different thresholds for assigning a peptide to be either binding or non-binding. In the first method, we use the raw affinity threshold of 500 nM, where strongly binding peptide will have an IC50 of less than 50 nM, whilst a weakly binding peptide will be between 50 and 500 nM. Peptides with an IC50 of greater than 500 nM will be assumed not to bind to the MHC allele. In the second method we classified all those peptides in the top 1% of all HIV peptides to be binders, meaning every allele binds the same number of peptides. We therefore carried out two sets of simulations in MATLAB. In the first set of simulations we simulated only those peptides with an IC50 less than or equal to the threshold of 500 nM. For the second set of simulations, in order to be able to compare the cell surface abundance of the peptides presented by each allele we required that the IC50 values between the alleles also be comparable, therefore we had to rescale the IC50 values according to the method in Ref. [159] and used in similar studies to this[143]. We acquired the predicted IC50 values for the peptides from the *Mycobacterium Tuberculosis* proteome to obtain a dataset of over $500,000$ partially overlapping natural peptides. For each allele studied here we obtained three separate datasets for the 9mers, 10mers and 11mers. We then combined the three datasets and took the top 1% of binders as the IC50 threshold for binding peptides for each allele. The rescaling method described in Ref. [159] normalises each IC50 by dividing by the threshold IC50 value. However, for our purposes we require a rescaled IC50 value that is still in units of nM. Therefore, we arbitrarily chose one allele as the reference allele and then rescaled the predicted IC50 values relative to that allele. The reference allele was chosen to be HLA-B*58:01, and its threshold affinity as determined using the method described above is denoted I_{B58}. When rescaling the predicted IC50 values for say HLA-B*57:01, we would multiply the IC50 value by the ratio of the threshold of B*58:01 to the threshold of B*57:01. Therefore, for allele a, the rescaled IC50 values are calculated as follows: $IC50_a^R = IC50_a * (I_{B58}/I_a)$, where $IC50_a^R$ is the rescaled IC50 of allele a, $IC50_a$ is the

original IC50 and I_a is the rescale threshold of allele a. The results of the simulations for method 1 are shown in Figures 5.10 and 5.11 and the results for the simulations for method 2 are shown in Figures 5.12 and 5.13. It can be assumed that the IC50 is approximately equal to the peptides dissociation constant from the MHC allele[160, 161, 162, 163], K_d, where $K_d = u_i/b_P$ where b_P is the peptide MHC binding rate. The binding rate was kept constant for each peptide, as experimental evidence shows that there is much less variation in HIV peptide binding rates than in the peptide off-rates[164] and the binding rate was set to $219\text{M}^{-1}\text{s}^{-1}$[165]. The peptide off-rate can then be determined for each HIV-1 peptide when binding to either a controlling or non-controlling allele.

5.3.4 Protein degradation and peptide cleavage in the cytoplasm

To be able to combine the HIV-1 intracellular kinetics model with the peptide filtering model[166] a few extra steps are required, which are not included in either of these other models, namely the kinetics of production of peptides from the HIV-1 proteins and the transport of these peptides to the ER. These steps are described in the following equation:

$$\frac{d[P_i]}{dt} = p_{i,j} \cdot k_{deg}^{Prot_j}[Prot_j] - g_i[P_i]_{cyt} - d_{P_{i,C}}[P_i]_{cyt}, \tag{5.17}$$

where P_i is the peptide of sequence i, cleaved from protein $Prot_j$. The production rate of peptide P_i is written as the product of the probability the peptide will be cleaved from the protein, $p_{i,j}$ and the degradation rate of the protein, $k_{deg}^{Prot_j}$. Peptide P_i can be degraded in the cytoplasm with rate constant $d_{P_{i,C}}$ and can also be transported to the ER with rate constant g_i. This is how the virion model and HIV intracellular kinetics model are connected to the peptide filtering model; the term $g_i[P_i]_{cyt}$ acts as the supply rate of peptide P_i to the ER.

We approximated values for the proteasomal cleavage probability for each peptide in the same way as in Chapter 4 Section 4.2.2. Briefly, the IEDB MHC-I processing tool[109] predicts a cleavage score for each peptide, which is proportional to the logarithm of the amount of peptide generated from the cleavage of the peptides C-terminal. To convert this in to a probability we scaled these values down by a factor of 1000, to get them to be within the expected range, as it has been measured that the well known

peptide SIINFEKL, or an N-terminally extended version of it, is produced via degradation of the OVA $6-8\%$ of the time it degrades[110]. SIINFEKL is known to be a highly immunogenic peptide, and so it is likely that it is produced with a high probability during proteasomal cleavage. Therefore, we set the highest potential probability of proteasomal cleavage to be 10%, and used this as an upper bound when predicting the proteasomal cleavage probabilities for the peptides in the model.

With this upper bound in mind, we took the amino acid sequences of HIV-1 consensus clade C proteins and used the IEDB MHC-I processing tool[109] to predict the cleavage scores for each peptide. As previously mentioned, the score is proportional to the logarithm (base 10) of the amount of peptide generated, therefore we converted these scores in to relative abundances of each peptide. These relative abundances ranged from between $1 \leq x \leq 80$, and so to obtain relative proteasomal cleavage probabilities for each peptide that lay within the above mentioned upper bound of 10% we divided the scores by 1000. These scores are therefore not supposed to represent a prediction of the *actual* proteasomal cleavage probabilities of each peptide, but rather the relative probability of cleavage.

Lazaro et al[167] investigated the variability of the cytosolic degradation rates of HIV peptides, and how this influences T-cell response. They found that their half-lives are highly variable and sequence specific, and this variability does in fact have a significant affect upon the efficiency of T-cell recognition. They then constructed an algorithm to predict the probability that a certain peptide sequence will have a half-life of either, 5 s, less than 30 s or longer than 2 mins. To do this they identified motifs which increase stability and those which decrease it, and so the prediction algorithm outputs a stability score for a specific peptide sequence. We therefore used their peptide stability tool to predict the half-lives and thus the degradation rate, $d_{P_{i,C}}$. of the HIV peptides used in the simulations.

The IEDB MHC-I processing tool predicts how well a specific peptide will be transported in to the ER by predicting an IC50 value for the peptide binding to TAP. Converting this value in to a supply rate for each peptide is more difficult than for the cleavage probabilities. The degradation of peptides in the cytoplasm is a much more efficient process

than the transport of peptides to the ER[168], and so this helps us put an upper bound on the possible range of the supply rate. A half-life of 5 s is the average peptide half-life in the cytoplasm[169], and is the lower-bound of the peptide stability predictions[167]. Therefore, we decided to set the supply rate of each peptide to be 0.08 peptides/s as this is less than the lower-bound of the peptide degradation rate i.e. $g_i \leq d_{P_{i,C,min}}$, where $d_{P_{i,C,min}}$ is calculated from the half-live of 5 s.

5.3.5 Self-peptides

Viral peptides compete not only with each other but also with peptides derived from the host's self-proteins or self-peptides for MHC class I binding and presentation. As self-peptides originate from native host proteins, they should not initiate a T-cell response due to negative selection of self-reactive T-cells. Just like the viral peptides, these self-peptides will have a range of proteasomal cleavage probabilities, ER supply rates and MHC class I binding rates. However, there are too many self-peptides to represent explicitly in the model. Therefore, to model the impact of the competition of these self-peptides we represented the self-peptides by four additional peptides in the simulation. These four peptides were given a range of unbinding and supply rates (see Table 5.2 for parameters). Each different MHC allele will only bind a small subsection of the ER peptidome with high affinity, and so the self-peptides with a medium unbinding rate (1×10^{-3} s^{-1}) were assumed to make up the majority of the peptides being transported in to the ER and were allocated a large fraction of the total supply rate. TAP binds between $2-5$ peptides per second, and there are approximately $10,000$ copies of TAP per cell[169]. We assigned the total rate of TAP transport of self-peptides to be the lower end of this range ($20,000$ peptides per second) to maximise viral peptide presentation. The kinetics

Parameter	**Unbinding rate** (s^{-1})	**Fraction of total supply**
Self-peptide-MHC unbinding very high	1×10^{-2}	0.05
Self-peptide-MHC unbinding high	1×10^{-3}	98.5
Self-peptide-MHC unbinding medium	1×10^{-4}	0.05
Self-peptide-MHC unbinding low	1×10^{-5}	0.05

Table 5.2: Self-peptide parameters. The values of the four representative self-peptide parameters used in the model.

of the self-peptides are given by are identical to those described for peptides in Equations

2.36 and 2.39-2.42 and 5.17. In total for each peptide in the system there are 5 equations describing them alone or in complex in both the cytoplasm and ER, so for n viral peptides and n_{self} self-peptides there are $5(n+n_{self})$ equations in the system. Therefore when simulating peptides derived from HIV proteins produced via host synthesis, there are $5(n+n_{self})$ equations in addition to the equations describing the HIV intracellular kinetics as described in Equations 5.1-5.16.

5.4 Results of the Mechanistic Model of HIV-1 Intracellular Kinetics and Antigen Presentation

5.4.1 A sensitivity analysis reveals the importance of different parameters changes in time

We performed a sensitivity analysis on a small subset of the model parameters which are peptide specific and influence cell surface abundance, using the SUNDIALS CVODES[170] forward sensitivity analysis (FSA) in MATLAB. The sensitivity of a system of nonlinear first order ODEs $\dot{x} = f(t,x,\theta)$ with respect to the kth parameter θ_k is computed as the partial differential of that function with respect to the parameter:

$$\dot{s_i} = \frac{d}{dt}(\frac{\partial x_i}{\partial \theta_k}) = \sum_{m=1}^{x_{dim}} (\frac{\partial \dot{x_i}}{\partial x_m}\frac{\partial x_m}{\partial \theta_k}) + \frac{\partial \dot{x_i}}{\partial \theta_k} \tag{5.18}$$

The CVODES FSA approximates $\dot{s_i}$ by a centred difference quotient. Both the sensitivities and ODE systems are solved simultaneously, to provide the time dependent parameter sensitivity. The sensitivity of the cell surface abundance of a single peptide, denoted here as MeP - competing against a group of self-peptides - was determined for the following set of parameters for each HIV protein: the peptide-MHC unbinding rate, u, the peptide-MHC binding rate, b_P, peptide-MHC-tapasin binding rate c_P, the peptide supply rate from cytoplasm to ER, g, the proteasomal cleavage probability of that peptide, ps, and the synthesis and degradation rates of the protein, f_j and k_j, respectively. The sensitivity coefficients s_i are normalised by the ratio of the baseline parameter value to the baseline system output i.e. $\theta_{k_0}/x_{i_0}(t)$ to remove the effects of units. Therefore, a normalised sensitivity coefficient of 1 indicates a positive linear dependency of the peptide cell sur-

face abundance upon that parameter, whilst a value of -2 indicates an inverse quadratic dependency.

For most proteins there was in general a linear dependency upon protein synthesis f_j over time (Figure 5.8A). However, the dependence of MeP on protein degradation (k_j) for each protein was sub-linear over time (Figure 5.8B), meaning cell surface abundance is in general less sensitive to changes in the protein degradation rate than the synthesis rate. A transiently non-linear dependency on protein synthesis rates were observed for the presentation of Tat and Rev peptides which may be due to the effects of nuclear shuttling. The presentation of the Gag peptides however, showed a sub-linear dependency on protein synthesis f_i, possibly due to the rapid accumulation of Gag epitopes in the cytoplasm resulting in the abundance instead being limited by ER translocation (Figure 5.8A).

For each protein there was a positive linear dependency of cell surface abundance on the probability of proteasomal cleavage ps, which remained constant over time, and a positive sublinear dependency on ER supply rate g which also remained constant over time. Comparing the sensitivity coefficients of the peptide-MHC binding rate b_P and the peptide-MHC-tapasin binding rate c_P reveals that the cell surface abundance has an almost linear dependency on c_P but the dependency on b_P is highly sublinear. Therefore, the peptide cell surface abundance is more highly influenced by the peptide binding rate to MHC-tapasin complexes than to empty MHC.

The sensitivity coefficient of peptide unbinding rate u becomes steadily more negative as time progresses, approaching an inverse quadratic dependency over time. Indeed, by 72 hours post infection the magnitude of the sensitivity coefficient for the unbinding rate is the highest out of all the parameters considered, suggesting that prolonged cell surface presentation depends highly upon the value of the unbinding rate, where lower values are more likely to be presented for longer. The sensitivity analysis highlights the trade-off occurring between the different parameters and their influence upon cell surface abundance changes with time.

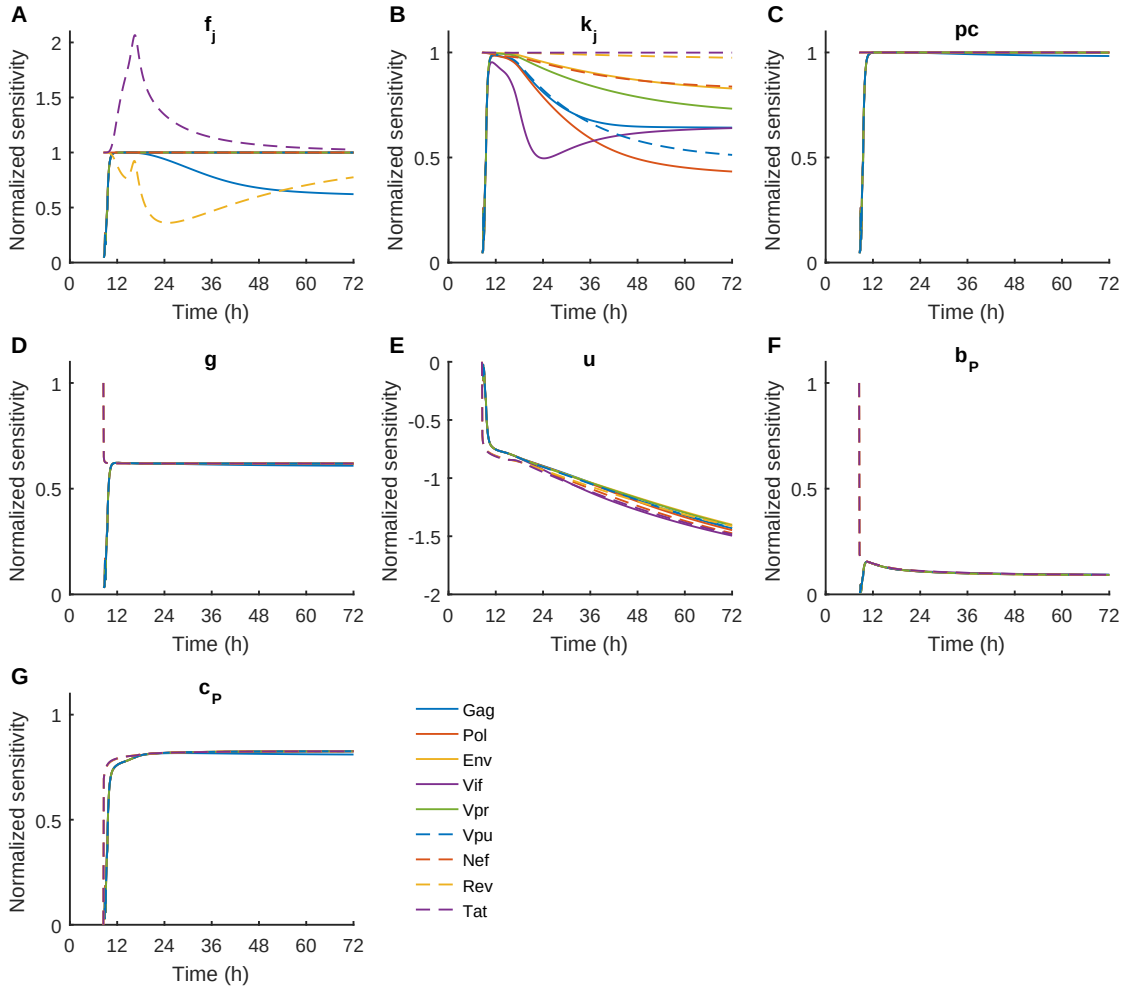


Figure 5.8: Sensitivity analysis of the combined model. We calculated the sensitivity of the cell surface presentation of optimal epitopes to seven of the model parameters: probability of protein translation f_j (j denotes the protein), cytoplasmic degradation k_j, proteasomal cleavage probability pc, supply rate to the ER g, peptide-MHCI unbinding rate u, peptide-MHCI binding rate b, and the peptide-MHC-tapasin binding rate c_P. The sensitivities were calculated using the CVODES module of the SUNDIALs package [171], then normalised.

5.4.2 An *efficient* Gag peptides dominate at the cell surface

We modelled the HIV-1 replication kinetics with deterministic ODEs, as HIV proteins are synthesised to high abundance in the cytoplasm following reverse transcription. The Gag protein had the highest cytoplasmic abundance in our simulations of HIV protein kinetics using the combined model described above, in keeping with the existing models of HIV kinetics. Experimental evidence suggests that the Gag:Pol ratio in the virion of 20 : 1 is maintained in the cytoplasm[172]. We then considered the presentation of an *efficient* peptide for each HIV protein, to determine which will produce the most abundant peptides

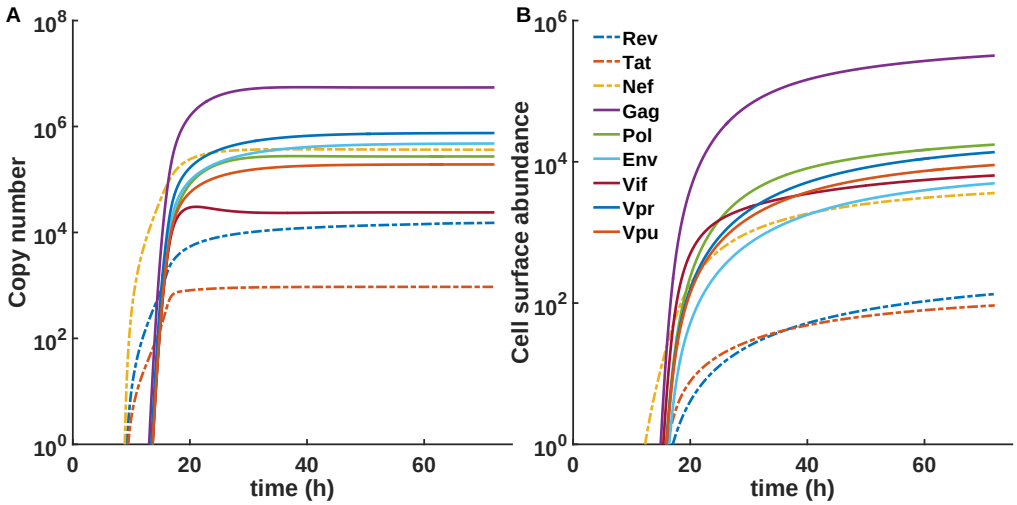


Figure 5.9: Example simulation of the combined model of HIV infection and peptide-MHC-I presentation. A) Simulated levels of HIV-1 proteins produced during replication (calculated deterministically). The complete model that produces all HIV-1 proteins is a combination of three existing models Kim & Yin[117, 173], Reddy & Yin[136], and Wang & LuHua[145]. B) Simulated cell surface abundance of *efficient* peptides derived from each protein, considered to have $u_i = 10^{-5}$ s^{-1}, a proteasomal cleavage $pc_{i,j} = 0.1$, and a fast supply rate $g_i = 0.08$ peptides s^{-1} to the ER.

on the cell surface if the peptide specific parameters are highly favourable and the same for each peptide so that only protein dynamics influence the cell surface abundance. We defined this efficient peptide as having a high affinity for MHC-I and so a low unbinding rate ($u = 1 \times 10^{-5}$ s^{-1}), a high proteasome cleavage probability ($pc_{i,j} = 0.1$), and a high supply rate to the ER ($g_i = 0.08$ peptides s^{-1}). The regulatory proteins Rev, Tat and Nef are the earliest proteins synthesised and are the first to appear in the cell cytoplasm at around 9 hours post infection. Peptides derived from these regulatory proteins are frequently targets of CTL response, and so may be good targets for a HIV-1 vaccines[174], before the down-regulation of MHC-I by Nef. Rev and Tat are continuously shuttling between the cytoplasm and the nucleus to regulate HIV mRNA nuclear export, and viral genome translation respectively[117]. The model predicts that this rapid shuttling results in slower cytoplasmic accumulation of Rev and Tat proteins than Nef, and even at late times when they have reached steady state the cytoplasmic abundance of these proteins is much lower than that of all other HIV-1 proteins (Figure 5.9 A). Therefore, out of the three early HIV proteins, the model predicts that only the Nef optimal epitope will be presented significantly early, around 12 hours post-infection. The optimal epitopes

from Rev and Tat however do not appear on the cell surface in substantial abundance until after the optimal epitopes from the later proteins, which appear at around 15-16 hours post infection (Figure 5.9 B). This suggests that any benefit that would be gained by targeting epitopes from the early HIV proteins before MHC-I down-regulation would only be applicable in the case of Nef epitopes and not Rev or Tat.

The Gag peptide dominates cell surface abundance, due to its very high abundance in the cytoplasm, with a Gag:Pol ratio of 18 : 1, which is similar to the ratio in the presentation of two HLA-A2 restricted peptides, gag 77-85 (SLYNTVATL) and pol RT 476-484 (ILKEPVHGV) of around 30:1[175]. The model predicts a Gag:Vpr ratio of 23 : 1 and a Gag:Env ratio of 64 : 1. The authors further suggest that this is consistent with the ratios of Gag and Pol in the virion and cytoplasm, around 20:1. However, unlike Gag, the ranking of the cell surface abundance for the optimal epitopes deriving from the other proteins does always follow the ranking of their cytoplasmic abundances. The second most abundant in the cytoplasm is Vpr, but the Pol peptide is the second most abundant on the cell surface, whilst Env is the third most abundant protein in the cytoplasm, however, its epitope has only the sixth highest cell surface abundance. This highlights the important trade-off between the rates of protein synthesis and degradation in determining peptide cell surface presentation: Vpr has a long half-life and so is very stable[149], therefore although it has a high cytoplasmic abundance, it degrades slowly, producing few peptides per unit time than faster degrading proteins such as Pol. A similar argument can be applied when considering the discrepancies in the ranking of the Env protein and its optimal epitope. Therefore, when trying to predict the cell surface abundance of an epitope the abundance and half-life of the protein from which it derives must also be considered.

5.4.3 HIV-1 intracellular kinetics and viral peptidome: LTNP vs fast progressors

Using the IEDB MHC-I binding tool to predict the unbinding rates and cleavage parameters for each peptide, we simulated deterministically the presentation of HIV-1 peptides by a set of HLA alleles associated with LTNP and a set associated with fast progression. The purpose behind this was to see if the simulations could explain why the presentation of Gag epitopes is correlated with control of HIV and to see if there are any obvious

differences in the presentation of HIV peptides by controlling alleles compared to fast progressors.

We simulated the cell surface presentation for four alleles associated with fast progression: HLA-B*18:01, -B*55:01, -B*07:02 and -B*35:03, and four alleles associated with long term non-progression: HLA-B*58:01, -B*57:01, -B*27:05 and -B*44:03. As described earlier we carried out two sets of simulations using different methods to define the threshold values for whether or not a peptide is binding or non-binding, and so will discuss the results separately.

Method 1: results We simulated the presentation of peptides of length $8-15$ amino acids long up to 72 hrs post infection and analysed the top 12 presented peptides at $16, 24$ and 72 hours. These simulations were carried out deterministically in MATLAB due to the large copy number of HIV proteins produced during HIV replication. At 16 hrs post infection HLA-B*58:01 and -B*57:01 presented majority Gag peptides on the cell surface with HLA-B*58:01 also presenting two Nef peptides, whilst HLA-B*57:01 only presented one Nef peptide. At 24 and 72 hours post infection both these alleles only presented Gag peptides in the top 12. The other two controlling alleles HLA-B*27:05 and -B*44:03 presented fewer Gag peptides at 16 hours post infection than HLA-B*57:01 and -B*58:01, and also presented Nef and Vif peptides. By 24 and 72 hours post infection the number of Gag peptides presented by these alleles increases with HLA-B*27:05 presenting majority Gag and HLA-B*44:03 presenting a mixture of Gag Pol and Vif peptides. In general using this method the controlling alleles are predicted to present largely Gag peptides in high abundance at all time points post infection.

At 16 hours post infection the non-controlling alleles HLA-B*07:02 and -B*18:01 present a mixture of Gag and Nef peptides in the top 12, just as the controlling alleles did. HLA-B*55:01 and -B*35:03, however, only presented one peptide at 16 hours post infection, from Gag and Nef respectively at very low abundance. At 24 hours post infection HLA-B*07:02 presents all Gag in the top 12, whilst HLA-B*18:01 presents a mixture of Gag and Pol peptides. HLA-B*18:01 continues to present a mixture of Gag and Pol peptides in the top 12 at 72 hours post infection, whilst three of the Gag peptides presented by HLA-B*07:02 at 24 hours post infection are replaced by Vpr, Pol and Nef.

The top 2 most abundant peptides presented by HLA-B*55:01 at 24 hours post infection are Gag peptides, however the rest of the top 12 is made up of a mixture of Vpr, Pol and Nef peptides, with a similar distribution at 72 hours post infection. HLA-B*35:03 does not present any Gag peptides in the top 12 at 24 or 72 hours post infection, but presents a mixture of Nef, Pol and Vpr peptides.

The simulations predicted that controlling alleles present a wide range of Gag epitopes at high abundance, many of which are known epitopes (Figure 5.10). For example, the simulations predict the well known B57/B58 p24 Gag epitopes KF11 (KAFSPEVIPMF) and TW10 (TSTLQEQIAW) are presented in the top 12 by HLA-B*58:01 and -B*57:01 as early as 16 hours post infection. Similarly, Gag epitope KK10 (KRWIILGLNK) is associated with control when presented by HLA-B*27:05, and is predicted to be presented in high abundance on the cell surface by 24 hours post infection. Furthermore, Gag AW11 (AEQATQDVKNW) is a well known and well defined epitope of HLA-B*44:03, and appears in the top 12 of this allele at 16 hours post infection. For the non-controlling allele HLA-B*35:03, the simulations predict that the peptides from Nef and Pol dominate the cell surface abundance, whilst for HLA-B*55:01, only one high abundance Gag peptide is presented, followed by peptides deriving from Pol. Both HLA-B*35:03 and -B*55:01 are predicted to present peptides at an overall lower abundance than the controlling alleles. The non-controlling allele HLA-B*18:01, however, presents a wide range of Gag peptides at a similar abundance to the controlling alleles, and with similar peptide sequences as HLA-B*44:03 dominating at the cell surface. However, HLA-B*18:01 does not present the HLA-B*44:03 epitope AW11, which as mentioned earlier is associated with control of HIV progression.

The results suggest that a wide distribution of high abundance HIV-1 Gag peptides (specifically Gag p24) presented on the cell surface may correlate with long term non-progression. Gag p24 is a highly conserved sequence in the HIV-1 genome, and mutations in this region negatively impact viral fitness.

Whilst HLA-B*58:01 and -B*57:01 presents HIV peptides in high abundance, their epitopes also experience high rates of mutation, the rate of which correlate with the CTL epitope-targeting frequencies. The epitope TW10 is the most rapidly escaping and most

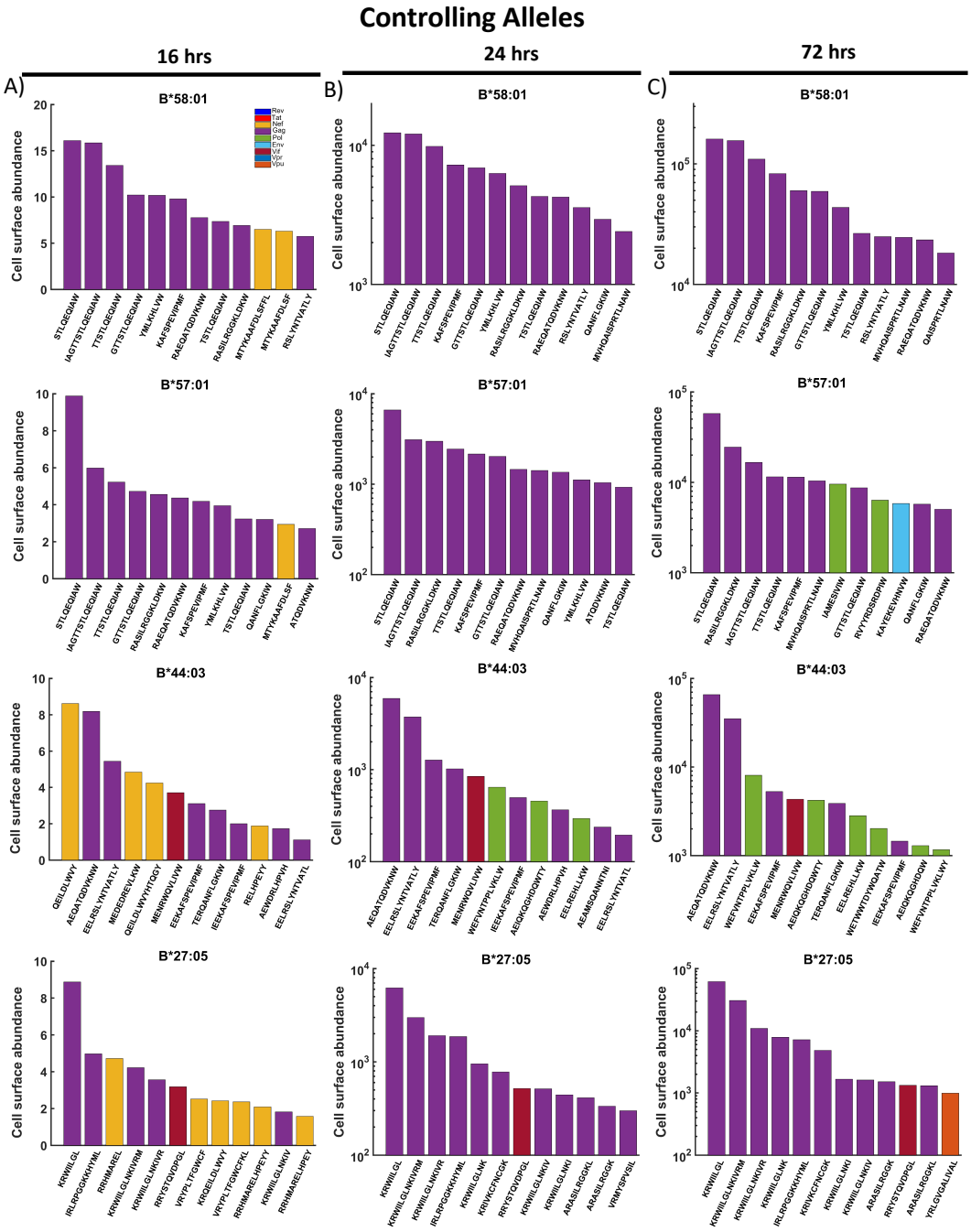


Figure 5.10: Method 1: Controlling alleles sustain Gag peptide presentation in high abundance throughout replication.The combined model was used to predict the cell surface abundance of HIV-1 peptides controlling alleles (HLA-B*58:01, -B*57:01, -B*44:03 and -B*27:05) over time. The top 12 most abundant peptides at A) 16, B) 24 and C) 72 hours post-infection are shown, with bar colours indicating the originating protein.

highly targeted B*57/B*58 restricted epitope[176]. Therefore whilst B*57 and B*58 are able to present in high abundance, their epitopes are also frequently subjected to escape mutations which may undermine its protective effect. However, it is known that the T242N mutation in TW10 leads to diminished viral replication capacity[129] suggesting that despite the high mutation rate, if some of these mutations are disadvantageous for the virus this may still result in control of viral progression. For HLA-B*27:05, the KK10 epitope (Figure 5.10) is the most frequently targeted epitope early in infection (out of all the HIV-1 epitopes studied in [176]), however its escape mutation rate is relatively low compared to other highly targeted epitopes. Furthermore, its mutations are linked to reduced capsid assembly[125], which may be the reason for its low mutation rate. As HLA-B*44:03 and -B*18:01 both present a similar range of Gag peptides at high abundance, but only HLA-B*44:03 is associated with control, we suggest control may rely on small differences in the sequences of presented peptides. An investigation in to mutations in the Gag-B44-AW11 epitope during acute infection found no evidence of escape mutations within this sequence[177]. This may because mutations in this particular region of p24 where AW11 is found impact viral fitness so negatively that mutated sequences quickly disappear from the population.

The conservation of the Gag sequence alone, however, cannot explain the association of Gag with control of HIV, as Env is also a highly conserved sequence of the HIV genome[178]. Overall these simulations suggest that Gag epitopes are associated with control of HIV because they dominate cell surface presentation, due to high turnover of Gag in the cytoplasm, i.e. its large synthesis rate and relatively high degradation rate compared to other highly abundant proteins such as Vpr and Env.

However, as mentioned earlier, the threshold of 500 nM affinity to separate binders from non-binders, is unreliable because the predicted IC50 values for peptides binding to different alleles are not comparable with one another. Therefore, whilst the results from method 1 produce some interesting findings, the method used to produce them is flawed and so the conclusions drawn above about the differences between controllers and non-controllers as predicted by the model cannot be substantiated. We therefore carried out a second round of simulations using method 2 as described above for more reliable

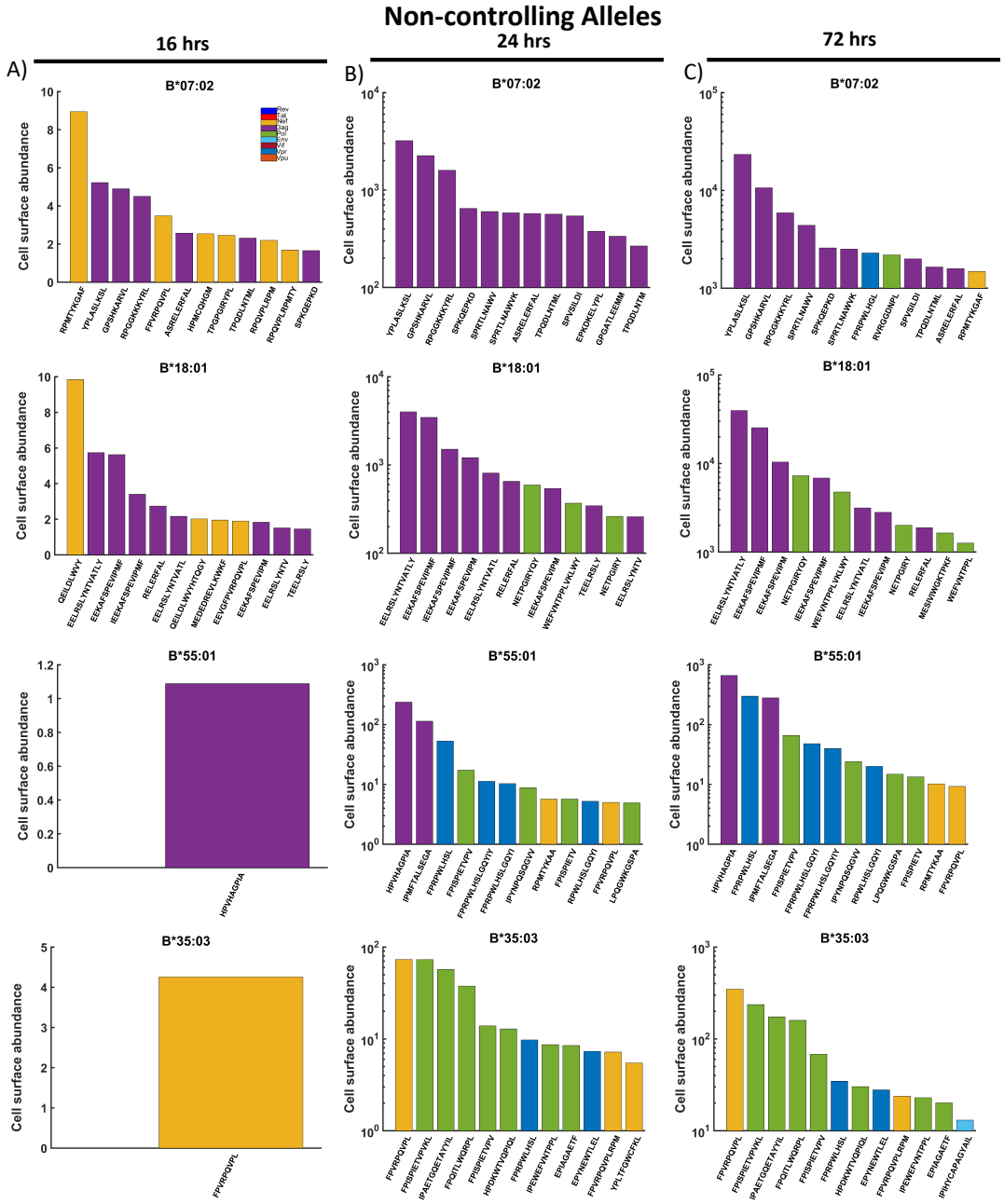


Figure 5.11: Method 1: Non-controlling alleles HLA-B*35:03 and -B*53:01 present only one peptide at 16 hours post infection and are unable to sustain high levels of Gag peptide presentation at later time points. The combined model was used to predict the cell surface abundance of HIV-1 peptides non-controlling alleles (HLA-B*07:02, -B*18:01, -B*55:01 and -B*35:03) over time. The top 12 most abundant peptides at A) 16, B) 24 and C) 72 hours post-infection are shown, with bar colours indicating the originating protein.

predictions.

Method2: results Using the IC50 rescaling method described in Section 5.3.3, we simulated the presentation of HIV peptides up to 72 hours post infection, and compared four controlling alleles HLA-B*58:01, -B*57:01, -B*27:05 and -B*44:03[123, 124, 125, 126, 138] and four non-controlling alleles, HLA-B*18:01, -B*35:03, -B*07:02 and -B*55:01[133, 139]. We analysed the top twelve most abundant peptides on the cell surface at three time points 16, 24 and 72 hours post infection, and determined which protein they originated from.

The results of the simulations using this method are very different than those using method 1. The cell surface abundance of the top 12 most abundant peptides at 16 hours post infection for the controlling and non-controlling alleles are shown in Figures 5.12 A and 5.13 A respectively. At this time point all alleles present a mixture of Gag peptides (purple bars) and Nef peptides (yellow bars) at highest abundance. At this time point there is very little difference between the controlling and non-controlling alleles, with the exception of HLA-B*27:05, which presents all Gag peptides in the top 12, with the exception of one Nef and one Vif peptide. The top 12 most abundant peptides at 24 hours post infection for controlling and non-controlling alleles are shown in Figures 5.12B and 5.13B respectively. At this time point all controlling alleles present majority Gag peptides in the top 12. Both HLA-B*57:01 and HLA-B*27:05 present exclusively Gag, whilst HLA-B*58:01 presents all Gag except the 12th most abundant peptide which originates from Pol. Similarly HLA-B*44:03 presents all Gag peptides except two Pol and one Vif peptide. At 24 hours post infection not one non-controlling allele presents all Gag peptides. HLA-B*55:01 presents the most Gag peptides, with 11 Gag and 1 Vpr, whilst HLA-B*07:02 presents 10 Gag and 2 Vpr. HLA-B*35:03 presents 9 Gag, one Vpr, one Nef and one Pol. However, the allele whose top 12 is most strikingly different from the controllers is HLA-B*18:01. The top two presented peptides by this allele at 24 hours post infection originate from Pol and Nef, with 8 Gag and 2 more Nef making up the remaining 10. The top 12 most abundant peptides at 72 hours post infection are shown in Figure 5.12 C for the controlling alleles and Figure 5.13 C for the non-controlling alleles. The controlling allele HLA-B*27:05 presents all Gag in the top 12, whilst the

other controlling alleles present fewer Gag peptides in the top 12 than they did at 24 hours. In the case of HLA-B*58:01 and -B*57:01, these Gag peptides have exclusively been replaced by Pol peptides. These Pol peptides have a very high affinity with the HLA alleles in question and so a low unbinding rate, and so this is an example of how as time progresses, the unbinding rate becomes more important than the abundance of the protein. The top three most abundant peptides presented by HLA-B*44:03 at 72 hours post infection are from Gag. However, the majority of the peptides presented by this allele in the top 12 are from Pol, with a single Env and Nef peptide also presented. The non-controlling allele HLA-B*18:01 presents majority Pol peptides at 72 hours post infection, with the Gag peptides that were in high abundance at earlier time points being replaced by high affinity Pol peptides. In fact this allele only presents two Gag peptides, both in the lower half of the top 12, whilst the top three peptides presented by this allele are Pol peptides. This suggests that this allele is unable to control the progression of HIV as it is unable to sustain high abundance presentation of Gag peptides. This also suggests that even if Pol peptides are presented in high abundance they will not induce a strong enough immune response to control HIV progression.

By 72 hours post infection, HLA-B*07:02 presents three Vpr peptides in the top 12, with the two Vpr peptides presented in the bottom two of the top 12 having increased in abundance and replaced Gag peptides at higher ranks. Both non-controlling alleles HLA-B*55:01 and -B*35:03 present fewer Gag peptides at 72 hours post infection than at 24 hours post infection, and both present one Gag and one Vpr peptide in the top 2 out of 12.

Therefore, one noticeable difference in general between the controlling and non-controlling alleles is that by 72 hours post infection, the controlling alleles seem to have a preference for presenting a mixture of Gag and Pol peptides in high abundance, whilst the non-controlling alleles have preference for presenting Gag and Vpr peptides in high abundance. This suggests that sustained Gag peptide presentation along with presentation of high affinity Pol peptides will be more effective at controlling HIV progression than a mixture of Gag and Vpr peptide presentation. We could also conclude that presentation of Gag peptides in high abundance alone is not sufficient to ensure control.

As with method 1, the simulations using method 2 also predicted the presentation

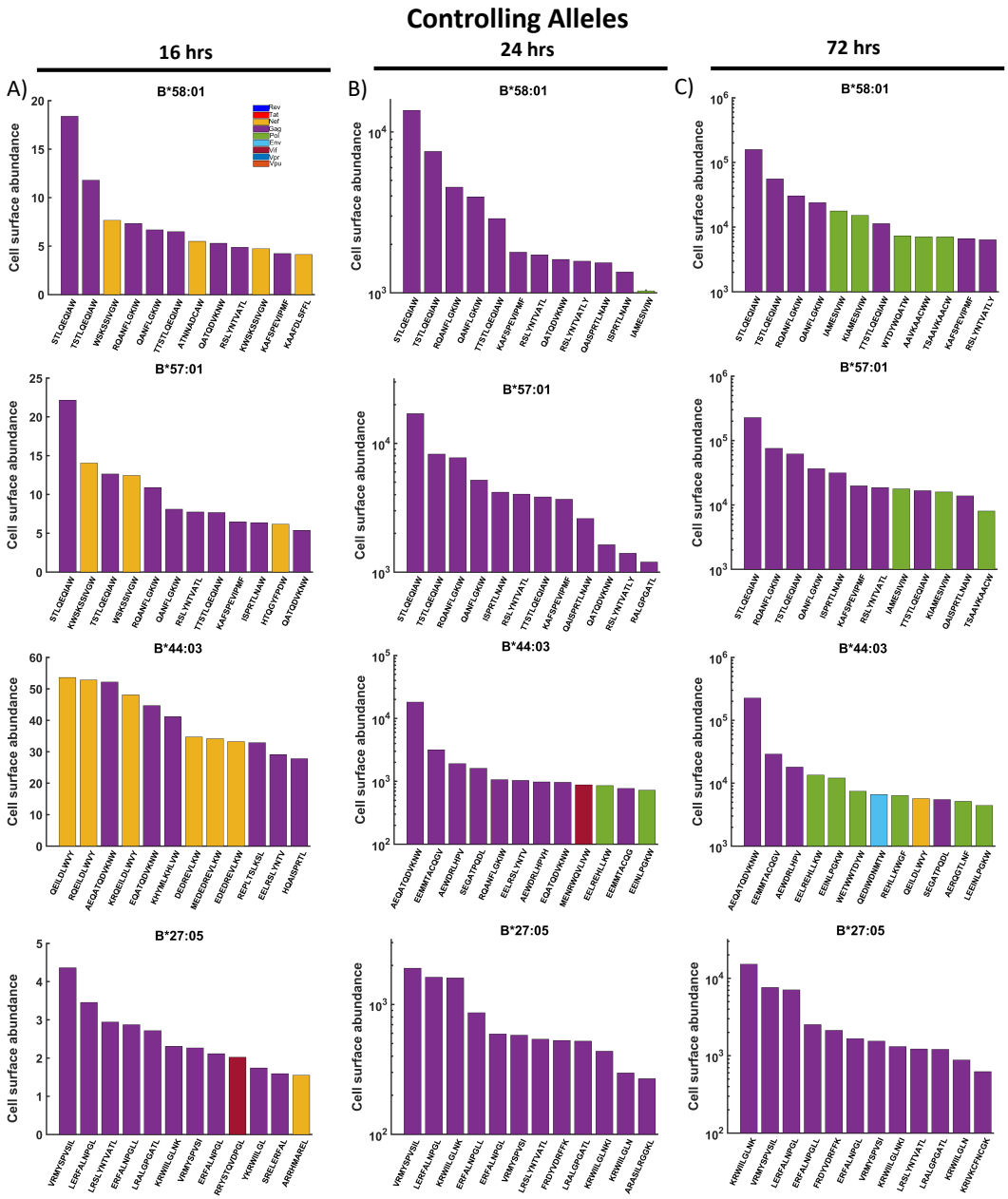


Figure 5.12: Method 2: Controlling alleles all demonstrate sustained Gag peptide presentation, and/or combined Gag and Pol peptide presentation at later times post infection The combined model was used to predict the cell surface abundance of HIV-1 peptides in controlling alleles (HLA-B*58:01, -B*57:01, -B*44:03 and -B*27:05) over time. The top 12 most abundant peptides at A) 16, B) 24 and C) 72 hours post-infection are shown, with bar colours indicating the originating protein. All controlling alleles presented several Gag peptides by 16 hours, with the number of Gag peptides increasing by 24 hours post-infection. The presentation of Gag peptides at high abundance is sustained up to 72 hours post-infection.

of known HIV-1 epitopes in the top 12 most abundant peptides. When comparing the outputs of the combined dynamic model (Figure 5.12 D) with those resulting from the static predictions of the IEDB processing tool (Figure 5.4), we found that the known epitopes rank higher in our combined model.

The model predicts that TW10 is the second most abundant peptide presented by HLA-B*58:01, and alternates between the second and third most abundant peptide presented by HLA-B*57:01 (Figure 5.12). However, the IEDB processing tool predicts TW10 to have only the 28th highest Total Score for peptides binding to HLA-B*58:01 and only the 34th highest score for HLA-B*57:01.

The known Gag p24 KF11 (KAFSPEVIPMF) epitope of HLA-B*58:01 and -B*57:01 is predicted to be the 11th most abundant HLA-B*58:01 peptide at 16 hours post-infection, then increases to the 5th most abundant peptide, before slipping down to 11th place by 72 hours post-infection (Figure 5.12). Similarly, KF11 increases its rank among the peptides presented on HLA-B*57:01, reaching 6th position by 72 hours post-infection. However, KF11 is only ranked 23rd and 16th by the IEDB Total Scores for HLA-B*58:01 and HLA-B*57:01 respectively.

Furthermore, the known B*58 and B*57 restricted Gag epitope ISPRTLNAW (IW9)[141] is the 11th most abundant HLA-B*58:01 peptide at 24 hours post-infection, but then displaced by Pol peptides at 72 hours post-infection. IW9 is consistently presented by HLA-B*57:01 at all three time points, being the 10th most abundant at 16 hours post infection, before increasing to 5th place and remaining there by 72 hours. IW9 has the 88th highest IEDB Total score for HLA-B*58:01 and the 54th highest total score for HLA-B*57:01.

Finally, the known HLA-B*27:05 restricted Gag epitope KK10 (KRWIILGLNK)[**?**] is the 6th most abundant peptide at 16 hours post-infection and the most abundant by 72 hours post-infection (Figure 5.12, bottom row), however it is only ranked 29th by the IEDB Total Score. Also, known HLA-B*44:03 restricted epitope Gag AW11 (AEQATQDVKNW) is the third most abundant peptide by 16 hours post-infection, but then reaches and remains the most abundant peptide from 24 hours onwards, however it has only the 13th highest predicted IEDB Total Score.

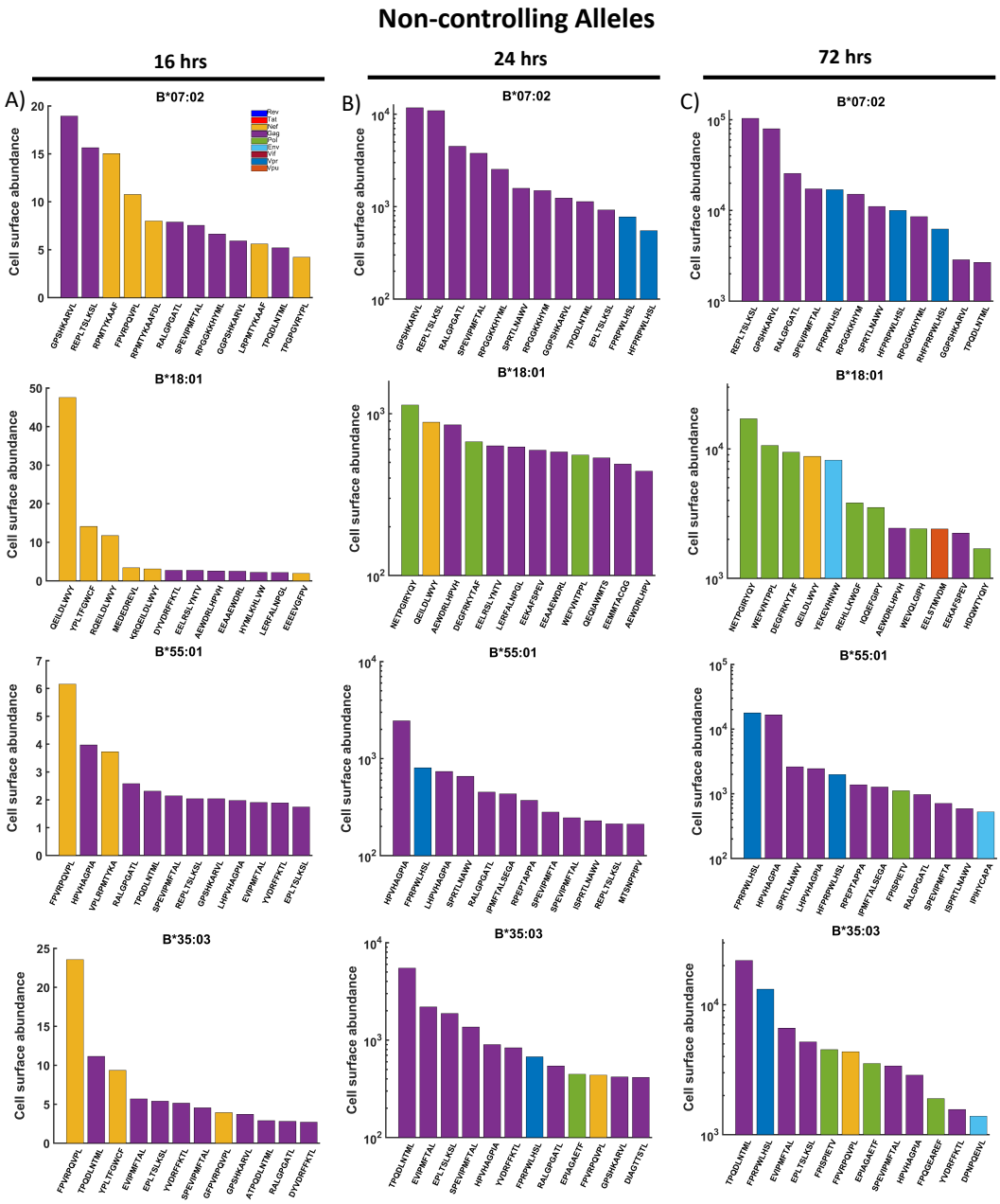
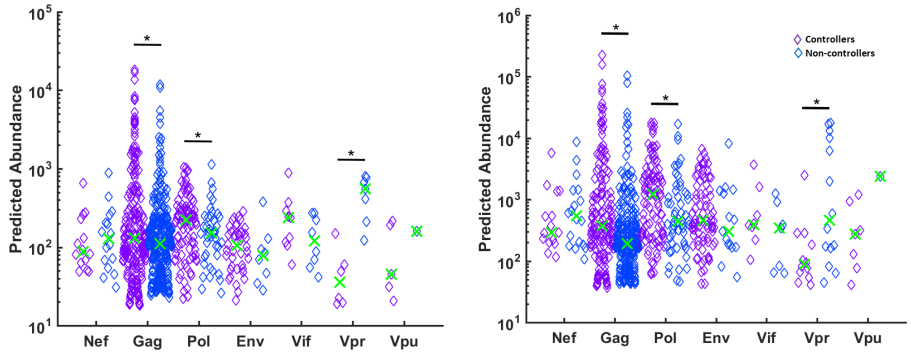


Figure 5.13: Method 2: Non controlling alleles are either unable to sustain high levels of Gag peptide presentation, or present a combination of Gag and Vpr peptides at later times post infection The combined model was used to predict the cell surface abundance of HIV-1 peptides non-controlling alleles (HLA-B*07:02, -B*18:01, -B*55:01 and -B*35:03) over time. The top 12 most abundant peptides at A) 16, B) 24 and C) 72 hours post-infection are shown, with bar colours indicating the originating protein.

The mixture of Gag and Pol peptides presented by controlling alleles may be due to the need for conservation of the Pol sequence compared to the need for conservation of the Vpr sequence. If escape mutations are able to reduce the impact of the T-cell response induced by Gag epitopes, further mutations in the HIV-1 genome, especially in the Pol region, may be disastrous for viral fitness, whereas mutations in the Vpr region may have a less adverse effect.

To investigate this further we carried out a similar analysis to that shown in Figure 5.5, but this time we analyse the top 1% most abundant peptides from each allele, grouped the data by controlling vs non-controlling, at both 24 (Figure 5.14 a) and 72 (Figure 5.14 b) hours post infection. We did not include 16 hours post infection due to the low number of peptides actually presented.

The median abundance of HIV peptides presented by controlling vs non-controlling alleles was analysed using a Wilcoxon rank sum test, which revealed controlling alleles peptides from from Gag (24hrs: $p = 0.0432$, 72hrs: $p = 3.67 \times 10^{-7}$) and Pol (24hrs: $p = 0.0435$, 72hrs: $p = 0.0135$) with statistically significant higher abundance than non-controlling alleles. Non-controlling alleles, on the other hand, were shown to present Vpr (24hrs: $p = 0.0023$, 72hrs: $p = 0.0288$) peptides with statistically significant higher abundance than controlling alleles. Interestingly, at 24 hours post infection the most significant difference between the controlling and non-controlling alleles is in the abundance of Vpr peptides, whilst at 72 hours post infection Vpr has the lowest significance, with the highest being associated with the abundance of Gag epitopes. The results of this analysis support our conclusions from considering the top 12 most abundant peptides from each allele, that non-controlling alleles have a preference for presenting Vpr peptides in high abundance, whereas controlling alleles preferentially present Pol and Gag. Furthermore, this also supports our position that whilst machine learning algorithms such as IEDB can predict differences in MHC binding affinities and proteasomal cleavage probabilities, using these tools alone it is not possible to predict differences in presentation as the protein dynamics are not taken in to account.



(a) 24 hours post infection **(b)** 72 hours post infection

Figure 5.14: Controlling alleles prefer to present Gag and Pol peptides. The predicted abundance of the top 1% of HIV epitopes from different HIV proteins were grouped by controlling and non-controlling alleles. The median abundance of the top 1% predicted HIV peptides at A) 24 hours and B) 72 hours post infection was then compared using a Wilcoxon rank-sum test. At both time points, a significantly higher median abundance was observed for Gag and Pol peptides in the controller group, while a significantly higher median abundance was observed for Vpr peptides in the non-controlling group.

5.5 Predicting HIV-1 Virion Antigen Presentation

The earlier viral antigen is presented on the cell surface, the sooner T-cells can detect it and destroy the infected cell, therefore early presentation is very important to successful immune control of a virus. The HIV-1 virion proteins provide the earliest source of viral peptides following virion entry to the cell. Identifying virion-derived epitopes that are presented before the down-regulation of MHC-I synthesis following the translation of the HIV-1 protein Nef may therefore be important in developing a successful HIV vaccine.

The HIV-1 intracellular kinetics antigen presentation model presented in the previous section predicts that peptides deriving from host-synthesised HIV-1 proteins are not presented until at the earliest 10 hours post-infection. However, Kloverpris et al. 2013[179] detected the two protective Gag epitopes KF11 and KK10, restricted by HLA-B*57:01 and -B*27:05 respectively, and the Pol KY9 HLA-B*27:05 restricted peptide on the surface of HIV-infected cells within 3 hours post-infection which could not have originated from *de novo* protein synthesis. These peptides were still observed when protein synthesis was blocked meaning they likely originated from the proteins that comprise the infecting virion(s).

We therefore sought to construct a mechanical model of the presentation of HIV

virion derived peptides. However, the number of copies of each HIV-1 protein within a virion is very low, therefore many of the assumptions of deterministic modelling break down, and so we had to simulate the virion model stochastically.

5.5.1 Methods

To model the virion protein dynamics we simply simulated the following reaction stochastically for each protein,

$$[Prot_j] \overset{k_{deg,Prot_j}}{\rightharpoonup} \emptyset, \tag{5.19}$$

where $Prot_j$ is the HIV protein j, and $k_{deg,Prot_j}$ is its degradation rate. At the beginning of the simulation, $t = 0$, the number of copies of each protein is set to the experimentally measured values found in a single virion. The proteins then degrade according to their degradation rate.

The most abundant protein in the virion is Gag, at 4900 copies per virion[135], and Vpr is found at a ratio of 7 : 1 to Gag, at around 700 copies per virion[135]. The Pol protein is found at a ratio 20 : 1 to Gag[172], with around 245 copies per virion. The copy numbers for the rest of the proteins in the virion are as follows: Vif: 101[145]; Env: 282[136]; Nef: 150[148]; Vpu: unknown; Tat: none; Rev: none.The degradation rates for each virion protein are the same as those given in Table 5.1.

Stochastic simulations are more computationally expensive than deterministic simulations, therefore we decided to simulate just one *efficient* peptide for each virion protein, similar to what was done in Figure 5.9 B.This efficient peptide has the same values of peptide specific parameters as before, a low unbinding rate ($u = 10^{-5}s^{-1}$), a high probability of proteasome cleavage ($p_{i,j} = 0.1$), and a fast rate of supply into the ER ($g_i = 0.08$ peptides s^{-1}). Whilst we cannot determine the presentation of specific individual peptide sequences in this way, we can determine how fast peptides can arrive at the surface in a near-optimal scenario, and thus establish an approximate upper bound. Equation 5.17 describes the kinetics of each efficient virion peptide.

Calculating the total probability of peptide presentation from each HIV-1 protein We simulate the presentation of an efficient HIV-1 epitope for each HIV-1 virion protein for the case where N virions enter a cell at one time, up to 9 hours post infection. We wished

to calculate the probability for each protein that the abundance of the efficient peptide is greater than or equal to 1. To do this we denote by $S_j(t)$ the surface presentation of the efficient peptide from protein j at time t. $P(S_j(t) \geq 1|N = i)$ denotes the conditional probability that the cell surface abundance $S_j(t)$ of peptide from protein j is greater than or equal to 1 when the number of virions N entering the cell is equal to i. The probability $P(N = i)$ denotes the probability that $N = i$ virions will enter the cell at one time.

We estimated $P(N = i)$ using the well-known multiplicity of infection principle [180], in which the number of infecting virions can be described by a Poisson distribution with mean 1, i.e. $N \sim \text{Poisson}(1)$. We chose to simulate the range $N = 1, 2, ..., 5$, which covers 99.9% of the probability mass for a Poisson with mean 1.

We estimate $P(S(t) \geq 1|N = i)$ by simulating the presentation of the efficient epitope from each HIV-1 protein for the range $N = 1, 2, ..., 5$ for 300 runs and determined the fraction of times the efficient peptide abundance is greater than or equal to 1 for each value of N. The total probability of presentation for each efficient epitope from each virion protein is found by summing over the probability of presentation greater than or equal to 1 for all values of N. Accordingly,

$$P(S_j(t) \geq 1) = \sum_{i=0}^{\infty} P(S(t) \geq 1|N = i).P(N = i) \tag{5.20}$$

By assuming that N is Poisson distributed with mean 1, we can easily combine simulations of the combined model for different numbers of virions.

5.5.2 Results: Gag-derived peptides are presented early following infection

We simulated stochastically the peptide presentation arising from degradation of HIV-1 virion proteins using Microsoft's Visual Genetic Engineering of living Cells (GEC)[181] software (freely available from `http://research.microsoft.com/gec`), which uses the domain-specific Language for Biochemical Systems (LBS) for specifying the reaction system. Due to the low protein copy number we simulated the model stochastically and averaged over 300 runs. The copy number of proteins contained in one virion, and their degradation over 9 hours is shown in Figure 5.15 A. Instead of simulating the

presentation of many peptides predicted by the IEDB prediction tools, we decided to simulate an optimal peptide for each virion protein. This optimal peptide has a very high affinity (and thus low unbinding rate) to MHC, a high probability of proteasome cleavage $p_{i,j}$ (in this case a 10% probability of being produced each time a protein is degraded), and a fast supply rate g_i. We did this to represent the best possible peptide that could be produced, and thus simulate the maximum possible presentation from virion proteins. We simulated this for a maximum of 5 HIV-1 virions entering a cell at one time. If we model the probability of a virion entering a cell using a Poisson distribution with mean $\mu = 1$ then $N = 5$ is at the tail end of that distribution and so represents the optimal possible situation which we could expect. The kinetics of the proteins contained in one virion over 9 hours post infection is shown in Figure 5.15 A. Each protein declines in abundance as it is degraded and converted into peptides. As N increases, the concentration of the proteins entering the cytoplasm increases, and so we would expect more peptides to be produced, thus increasing the probability of peptide cell surface abundance of greater than or equal to 1. For all values of N, the efficient Gag peptide is the most abundant on the cell surface and the total probability of Gag peptide presentation converges towards 0.63, equal to the probability that at least one virion infects the cell, i.e. $P(N > 0)$.

By calculating the total probability of presentation we predict that an efficient Gag peptide has a greater than 50% chance of being presented by 3 hours post-infection, reconciling the observations of Kloverpris et al. 2013[179]. Our model also predicted that an efficient Vpr peptide has a higher probability of presentation than an equivalently efficient Pol peptide. This may be because the Pol degradation rate we are using in this simulation is the same as the Gag-Pol polyprotein degradation rate. However, the Gag-Pol polyprotein is cleaved in to the enzymes integrase, reverse transcriptase and protease. The degradation of these smaller constituent proteins could be faster than that of the polyprotein, which will affect the timing and probability of presentation of the peptides cleaved from the Pol enzymes.

Therefore, the model predicts the probability that an efficient non-Gag peptide is presented following the entry of a single virion is low, however, if multiple virions enter this probability could increase and become immunologically relevant.

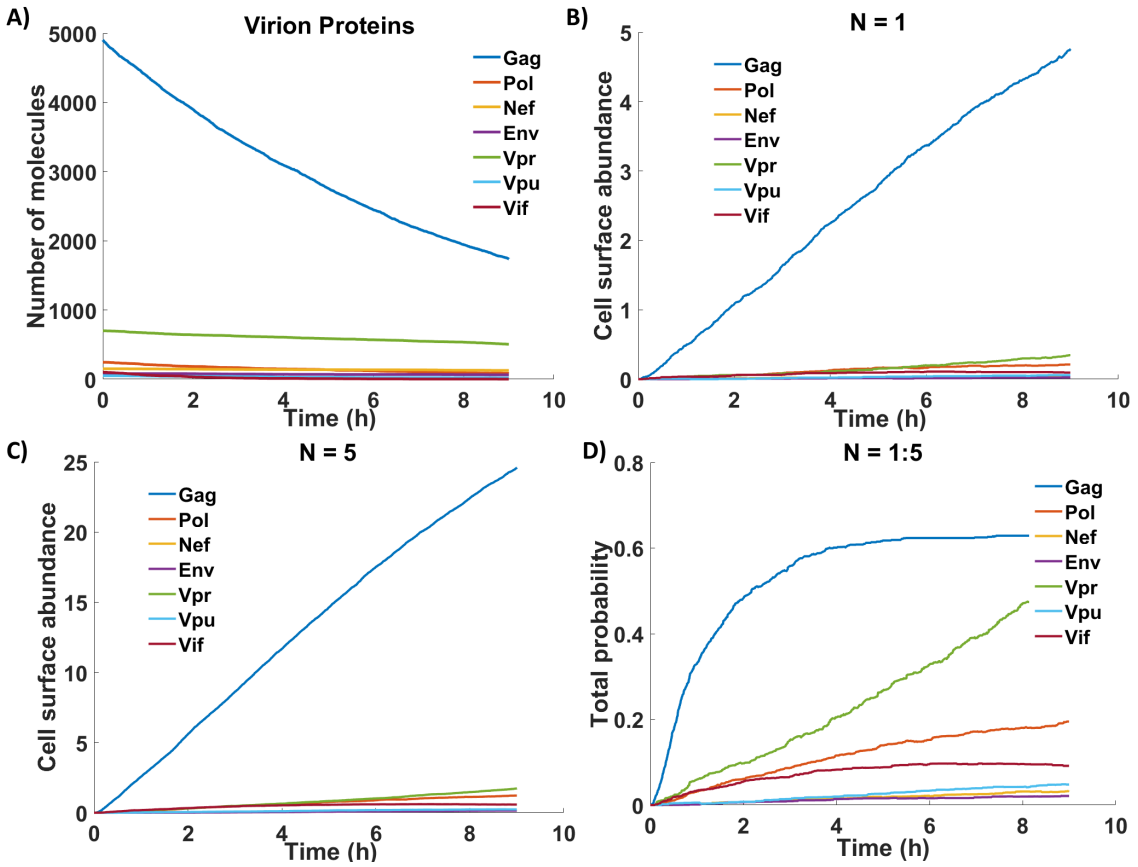


Figure 5.15: The virion model was simulated stochastically and averaged over $M = 300$ runs and the average HIV virion protein kinetics are shown in A). We simulated the presentation of a single, optimal peptide from each of the 7 HIV proteins contained within a virion, and calculated the conditional probability of presentation by multiplying $n(\geq 1)/M$ by the Poisson probability $P(N) = \lambda^N e^{-\lambda}/N!$, where the mean, $\lambda = 1$ and $N = 1,2,3,4,5$. Example simulations for $N = 1$ and $N = 5$, where N is the number of virions entering the cell are shown in B) and C) respectively. D) We calculated the total probability of presentation for each peptide by summing the conditional probability of presentation over N

5.6 Discussion

We sought to construct a model comprised of existing sub-component models that incorporates viral intracellular kinetics, protein sequence specificities and MHC-I affinity to produce dynamic predictions of resulting viral peptidome. We chose to study HIV-1 here, but the general model can be applied to any virus, provided its intracellular kinetics are known. These predictions may shed some light on the characteristics of HIV-1 peptide presentation by HLA alleles which control the spread of infection. We hoped to find some explanation for the association of Gag peptides and a strong T-cell response against HIV infection.

Initially we used the IEDB MHC-I processing tool to obtain predictions of a 'Total Score' for each peptide from each HIV-1 protein when binding to a set of controlling (HLA-B*58:01, -B*57:01, -B*27:05 and -B*44:03) and non-controlling (HLA-B*07:02, -B*18:01, -B*55:01 and -B*35:03) alleles. We used two different methods to select which set of peptides we expected to bind to the MHC-I allele in question. The first method used a IC50 cut-off of 500 nM, and the second method looked at the peptides whose Total Score was within the top 1% for all HIV-1 peptides. We will only discuss the results from the second method here as any conclusions drawn from the first method will be unreliable as it is not actually possible to compare IC50 scores between alleles.

From the IEDB predictions of the peptides Total Scores we would expect Pol or Env peptides to be associated with a strong T-cell response, as a large fraction of all predicted peptides from the entire HIV genome originate from these two proteins, and in general, the peptides from these two proteins are associated with higher Total Scores than those from Gag. Furthermore the most significant difference in median Total Score between the group of controlling alleles vs the group of non-controlling alleles was observed for Pol ($p = 4.8 \times 10^{-18}$) and Env ($p = 2.9 \times 10^{-24}$), whereas the Gag was associated with a lower significance ($p = 3.4 \times 10^{-6}$), where significance was determined using a Wilcoxon rank sum test. In fact, this analysis revealed that there was a significant difference in the Total Score distributions of *all* HIV-1 proteins when comparing controllers vs non-controllers. Therefore, the IEDB predictions alone cannot explain why Gag peptides are so often associated with HIV-1 control.

Gag is a highly conserved sequence in the HIV-1 genome, and mutations in this region negatively impact viral fitness, however, other highly conserved sequences such as Pol are not associated with a similar degree of immune response. By combining existing models of HIV-1 intracellular kinetics, with a model of peptide-MHC binding and presentation along with the IEDB prediction results, we were able to predict the cell surface abundance of HIV-1 peptides on a single cell following infection.

The IEDB Total Score predictions show that on average peptides from Pol and Env are the most immunogenic. However, in the simulations, Gag peptides dominate the cell surface peptidome for both the set of controlling and non-controlling alleles. This is due to

the large turnover of Gag in the cytoplasm during the replication cycle. If Env or Pol had similar kinetics to Gag it is likely that their peptides would out-compete Gag for binding to the controlling HLA alleles because they are predicted to have on average higher Total Scores.

By simulating peptide presentation by the groups of controlling and non-controlling alleles we were able to analyse the temporal distribution and cell surface abundance of HIV-1 peptides up to 72 hours post-infection. The peptidome at 16 hours post infection is much different than at later time points, with mainly Gag and Nef peptides being presented in low abundance. At 24 and 72 hours post-infection the top 12 presented peptides of the controlling alleles were in general a mixture of Gag and Pol peptides, whilst non-controlling alleles presented a mixture of Gag and Vpr. Upon conducting a Wilcoxon rank sum test at 24 and 72 hours post-infection we found that controlling alleles presented significantly more Gag and Pol peptides, whereas the non-controlling alleles presented significantly more Vpr peptides. We therefore conclude that a combination of Gag and Pol peptides in high abundance on the cell surface will result in greater control of HIV-1 infection that a combination of Gag and Vpr peptides.

By analysing the specific sequences of the top 12 most abundant peptides, we observed that many known Gag along with some known Pol eptiopes were presented by the controlling alleles. It is interesting to note that these known epitopes dominate at the cell surface in the combined model predictions, whereas using the IEDB predictions alone their Total Scores were not among the most immunogenic of all HIV-1 peptides.

We carried out similar simulations for peptide presentation from virion entry, using experimentally measured levels of HIV protein copy numbers within a virion. The model predicts the presence of at least one Gag peptide by 2 hours post infection. All peptides in the virion model had the same proteasomal cleavage, peptide supply, and MHC binding and unbinding rates, therefore the dominance of the Gag peptide can only be a result of the protein's high abundance within the HIV-1 virion.

These simulations highlight the importance of protein kinetics to MHC-peptide cell surface presentation. Peptides in the cytoplasm have very short half-lives ($\approx 5s$[169]) and so only 1% of them even survive to encounter TAP. Therefore, in order to enter the ER,

a peptide must be in very high abundance in the cytoplasm, requiring very high protein abundance. The simulations show that even if peptides are being produced from low abundance proteins (such as Env and Vif for example) with a very high probability of being cleaved from the protein and high affinities to the MHC allele, they will still be outcompeted for cell surface presentation by high concentration peptides of lower affinities. For example, the peptides with the highest Total Scores when binding to HLA-B*57:01 come from Env (Figure 5.4), however, there is not a single Env peptide present in the top twelve most abundant peptides presented by this allele at any time point (Figure 5.12). Indeed, at 24 hour post infection the top 12 peptides presented by HLA-B*57:01 are entirely derived from Gag, with three Pol peptides appearing at 72 hours post infection. However in the model, Pol degrades much faster meaning it will produce more peptides within a certain time frame than Env will. (Note: the degradation rate of Pol in the model was set to equal that of $1.1 * k_{Gag}$ because this value, along with the probability of GagPol mRNA translation $f_{GagPol} = 0.05$, and the budding rate k_{bud} results in a ratio of Gag:Pol in a single virion of $20:1$. The synthesis and degradation of Env are the values presented in [147]).

These simulations indicate the importance of considering protein kinetics when predicting epitope candidates. Being able to detect infection in the early stages before replication may be key to creating an effective HIV-1 vaccine, and this model suggests that the only way to do that is to target Gag epitopes, as the presentation of peptides from other HIV-1 proteins cannot be guaranteed. Therefore, tools for predicting potential T-cell epitopes from protein sequences alone, such as IEDB should not be taken at face value, but should be combined with knowledge of the protein kinetics in order to establish targets for peptide vaccines.

Chapter 6

General Conclusions

6.1 Introduction

The T-cell response to the presentation of foreign or mutated peptides is key to the immune system's ability to control infection from viruses and bacteria, as well as controlling the spread of cancer. In the case of HIV, there is a strong association between disease progression and T-cell responses against certain epitopes presented by certain HLA alleles. In the case of cancer, many patients develop endogenous anti-tumour T-cell responses, but these responses are non-protective[182, 183]. Furthermore, immunotherapy will work in some patients but not others.

Understanding why some T-cell responses are non-protective whilst others are protective is vital for understanding the immune system and designing effective therapies. Therefore, it is important to develop robust methods to identify T-cell epitopes, especially those which result in protective T-cell responses.

T-cell epitopes can be discovered using high-throughput experimental methods, such as enzyme linked immunospot (ELISPOT), and cytokine flow cytometry (CFC), however there are over 12,000 HLA alleles and each is highly promiscuous, therefore to scan the T-cell response to all possible pathogenic peptides for even one HLA allele would be highly time consuming and expensive. As a result, it is infeasible at present to perform full scans of all potential T-cell epitopes for complex viruses such as HIV[58].

Therefore, mathetmatical models which can predict the immunogenicity of a peptide can be very useful. As discussed in Section 2.2, machine learning tools which predict

immune parameters such as the IC50 of the peptide-MHC binding[61, 62] can be used to quickly screen the pathogenic sequences that might be possible T-cell epitopes and reduce the number of possible targets to consider experimentally. However, the correlation between the predicted affinity and cell surface abundance is in general quite low[90] (where a lower IC50 means a higher affinity and a better binder, which should result in higher cell surface abundance). High peptide cell surface abundance has been found to correlate with immunogencity[59, 99], therefore the development of a model which is able to predict peptide cell surface abundance is of high importance, as it could help narrow down the candidate epitopes for T-cell related therapies.

Dalchau et al. 2011[22] constructed a model of peptide binding and presentation by MHC-I alleles, as described in Section 2.6. The model quantifies the key peptide-specific properties of intracellular peptide processing, including peptide supply into the ER and peptide MHC-I unbinding. In this thesis we extend the model to include the impact of varying cytoplasmic abundance on peptide ER concentration, and we tested the model's ability to predict peptide cell surface abundance.

Firstly, we calibrated the model on an experimental data set of a target and competitor peptide with varying cytoplasmic concentrations, with and without the addition of IFN-γ (Figures 3.5 and 3.6). The model was in general able to capture the impact of changing competitor abundance upon the cell surface presentation of the target peptide and visa vesa, with the lowest NRMSE was associated with SSL surface data without the addition of IFN-γ. The model was also able to account for the decrease in competition between the two peptides when IFN-γ was added (Figures 3.5 and 3.6 panels B and D).

Next, we tested the calibrated model on competitor peptides with varying MHC-I unbinding rates. The model was able to largely predict the cell surface presentation of the target peptide SSL and the varying competitor peptides with different MHC-I unbinding rates (Figure 3.7). We measured how well the model predicted the data by calculating the normalised root mean square error (NRMSE) between the data and the model, where a low value indicates a good fit.

Whilst the model predicted the competition data well, we also calculated a simple competition metric, based upon the filter relation presented in [22], to determine if the

metric could approximate the output of the model for the cell surface abundance of SSL (Section 3.5.1, Equation 3.6). We calculated the Pearson's correlation coefficient between the competition metric and the cell surface abundance of SSL predicted by the model, with and without IFN-γ, for each variant of the competitor peptide ASN (Figures 3.9 and 3.10). We observed a high correlation between the metric and the model output for each simulated experiment, with correlation coefficients ranging between 0.74 and 0.86, meaning the competition metric approximates the steady state behaviour of the model well.

We also calculated a similar peptide competition metric using the peptide cytoplasmic data (Section 3.5.1, Equation 3.7), and compared the results with the SSL cell surface data calculating the Pearson's correlation coefficient as before for with and without IFN-γ (Figures 3.11 and 3.12). We observed that the correlation coefficient was much higher for the dataset without IFN-γ, with values ranging between 0.52-0.80, compared to a range of 0.36-0.57 for with IFN-γ. We concluded this was most likely due to the missing contribution of self-peptides which we were unable to quantify for the data metric (Equation 3.7), whereas this value was included in the simulation metric (Equation 3.6) using the calibrated self-peptide parameter values.

Having tested the peptide competition metric against experimental data of *two* peptides (three including the one representative self-peptide) competing, we then wanted to test how well the normalised filter relation (Equation 4.8) could approximate simulations of an entire peptidome of competing peptides (Chapter 4). To do this we used experimentally measured protein abundance and half-life and predicted the unbinding rates of each peptide from each protein using the IEDB MHC-I binding prediction tool. The normalised filter relation for peptide i is the ratio of the supply rate g_i to the square of the peptide off-rate u_i^2, divided by the sum of the ratio g_k/u_k^2 for each of the peptides k that peptide i is competing against. Here we approximate g_i as the product of the protein abundance, protein degradation and the probability of cleavage, as predicted by the IEDB MHC-I processing tool. We therefore considered the four parameters off-rate, protein abundance, protein degradation and proteasomal cleavage separately and found a weak correlation between these parameter values and the cell surface abundance of the peptide

(Figure 4.1 A,B,C and D). However, when combining these parameters in to the filter relation we observed a very strong correlation of 0.998, demonstrating that the filter relation applied to each peptide can be used to predict the relative cell surface abundance of an entire peptidome competing for the same MHC allele.

We then compared the performance of the normalised filter relation (Equation 4.8) and the raw filter relation (Equation 4.9), in predicting peptide cell surface abundance. We found that the raw filter relation performed just as well as the normalised filer relation. This is likely because for a very large number of peptides competing in the ER, the normalisation factor in Equation 4.8 will be the same for each peptide, as no one peptide will dominate. This suggests that if one wished to predict the cell surface abundance of a peptide when in competition with an entire peptidome, all one would require would be data quantifying the peptide supply rate to the ER and the peptide-MHC unbinding rate of only the peptide in question.

We also demonstrated that our model could be used to predict which neo-epitopes derived from mutated tumour proteins would be presented in highest abundance on the cell surface, and compared our results to those of Boegel et al. 2014 [112] who predicted the IEDB affinities of peptides deriving from mutated proteins and chose those peptides with the highest affinity to be the most immunogenic. Unsurprisingly, our set of abundant neo-epitopes was different from Boegel et al.'s, however we observed only a small overlap between the dataset we used to quantify the abundance and degradation of the HeLa cell proteins (Nagaraj et al. 2011[108]), and the Broad-Novartis Cancer Cell Line Encyclopaedia (CCLE)[111] dataset of HeLa cell mutations. Therefore, a larger set of HeLa cell protein abundance and degradation measurements is required to improve these predictions.

Next we wished to simulate a viral peptidome, however the temporal appearance of viral proteins is very important in determining the timing and abundance of viral epitopes. In this case the competition metric would be of no use unless it was calculated at many different time points. We chose to model HIV-1 peptide presentation and so we combined already existing mechanistic models of HIV-1 intracellular kinetics with the peptide filtering model, and used the IEDB p-MHC binding prediction tools to estimate

peptide-specific parameters.

When simulating a highly efficient peptide originating from each of the 9 HIV-1 proteins the model predicted that the Gag peptide would dominate at the cell surface (Figure 5.9 B). As discussed in Section 5.2 varying rates of progression for HIV-1 to AIDS are observed depending upon the HLA alleles expressed by the infected individual. Certain MHC-I alleles associated with long term control of HIV-1 are associated with a strong T-cell response against the epitopes they present, with the majority of the known T-cell HIV-1 epitopes originating from the Gag protein.

The results of the machine learning algorithms alone cannot explain why Gag is so immunogenic (Figures 5.4 and 5.5), as the Gag peptides do not have the best Total Scores and when comparing between the controlling (HLA-B*58:01, -B*57:01, -B*27:05, -B*44:03) and non-controlling group (HLA-B*07:02, -B*18:01, -B*53:01, -B*35:03) there is a significant difference in Total Scores for *all* HIV-1 proteins.

We simulated the presentation of HIV-1 peptides by the controlling and non-controlling alleles and found that at both 24 and 72 hours post infection the abundance of Gag and Pol peptides presented by the controlling group was significantly greater than the non-controlling group, whilst the abundance of Vpr peptides was significantly greater in the non-controlling group (Figures 5.12, 5.13, 5.14). We therefore conclude that control of HIV-1 requires high abundance of both Gag and Pol peptides, whilst a high abundance of Gag and Vpr is associated with fast progression.

Furthermore, the model predicts that the controlling alleles will present several known HIV-1 epitopes in the top 12 most abundant peptides (Figure 5.12), whereas the rank of these epitopes among the predicted IEDB Total Scores are much lower. This demonstrates that only predicting peptide specific parameters such as MHC affinity and proteasomal cleavage is insufficient and highlights the importance of considering protein kinetics when predicting possible T-cell epitopes.

We also simulated the presentation of HIV-1 peptides following virion entry in to the cytoplasm. The copy number of HIV-1 proteins in a single virion is very low and so the concentration of peptides resulting from the degradation of these proteins will be even lower. When dealing with such low concentrations the assumptions of deterministic

modelling no longer hold and so we simulated the presentation of an efficient peptide from each protein stochastically. The model predicts a 50% chance of an efficient Gag peptide being presented at 3 hours post infection, agreeing with experimental observations of Kloverpris et al 2013[179].

6.2 Further Work

This work has shown that the prediction of cell surface abundance of pMHC requires embedding peptide sequence based algorithms in models that also incorporate chemical rate equations describing protein kinetics and the known mechanisms of the antigen processing and presentation pathway, as machine learning predictions are static and so their power when used alone will always be limited. The aim of such a model is to provide reliable predictions of cell surface abundance and possible T-cell epitopes to narrow down the number of peptides experimentalists have to consider.

In this work we have focussed on putting together such a model for HIV-1 however this approach can be applied to any virus, provided the relevant data is available. In order to be able to construct such a model, the important parameters in the viral intracellular life-cycle are required, such as mRNA transcription and protein translation or, more simply, quantitation of viral protein abundance and degradation over time in the cytoplasm following virion entry and *de novo* synthesis. Once such data is acquired it can be combined with predictive machine learning algorithms and already existing models of the antigen processing and presentation pathway, such as was here done for HIV-1.

HIV has been extensively studied and so models of HIV-1 viral intracellular kinetics and additional parameters such as protein half-life already existed in the literature. However, this is not the case for the majority of viruses. Therefore, in order to progress in this field of research the acquisition of such data is required, and so biologists must begin to design experiments with the aim of furnishing such quantitative, predictive models. As has been demonstrated here, such models can be genome specific by considering only the immune molecules (such as MHC-I) expressed by a certain individual. Such models therefore have broad applications in the field of personalised medicine, where rapid simulations could be used to help decide upon treatment dependent upon the patients genome.

6.2.1 The impact of DRiPs

Up to this point we have not considered the contribution of Defective Ribosomal Products (DRiPs). As mentioned in Chapter 1, DRiPs are defined as "prematurely terminated polypeptides and misfolded polypeptides produced from translation of *bona fide* mRNAs in the proper reading frame"[15]. In other words DRiPs are generated by the degradation of proteins synthesised during the pioneer round of translation, proteins whose translation has been prematurely terminated, and proteins which misfold or fail to assemble into their multisubunit protein complexes. Yewdell[15] first proposed that peptides might derive from DRiPs in 1996 to explain the rapid presentation of viral T-cell epitopes. However, our simulations along with experimental evidence such as Kloverpris et al. 2013[179], demonstrate that early T-cell responses can result from the processing and presentation of proteins comprising the infecting virion(s).

However, it has been proposed that DRiPs make up a significant source of self and viral peptides[15, 16], and rapid degradation of these newly synthesised proteins may allow T-cell recognition of foreign or mutated peptides much earlier than if peptides derived from the degradation of native proteins alone. However, there is some disagreement over just how significant DRiPs as a source of peptide. Several studies have provided evidence that DRiPs contribute upwards of 30% of all degraded proteins[184, 185]. However, this would mean that a large amount of energy is essentially wasted on producing non-functional proteins[186].

To incorporate DRiPs in to, for example, the model of HIV-1 antigen presentation the rate of translation of DRiPs from the HIV-1 mRNA, and the degradation rate of the corresponding DRiPs would need to be experimentally determined or estimated. The production of peptides via proteasomal cleavage of these DRiPs could be modelled in the same way as was done in this thesis, and the resulting peptide pool could then be connected to the antigen processing model as described in Equation 5.17.

Even without experimental data characterising the synthesis and degradation of HIV-1 DRiPs, some simple predictive simulations could be carried out to determine what values of the synthesis and degradation would result in either some percentage increase in cell surface abundance, or an earlier presentation time.

6.2.2 Including the T-cell response

The T-cell response to a presented epitope is known to depend upon both the dose and the affinity between the TCR and the pMHC complex. The model we have presented in this work does not yet account for the affinity between the pMHC complex and the TCR and how this determines the T-cell response. Therefore, the next step in this area of research would be to combine the model of peptide cell surface presentation with existing models of T-cell signaling such as Lever etal.[187]. To make this peptide specific, the affinity between the pMHC and the TCR would need to be predicted. One way of doing this is to use molecular dynamics simulations. For a more detailed description of how this would work see Eccleston et al[188].

Appendix A

Abundance Project: LBS Code

```
// Simulate for some time in the past, add inputs at
//time 0.0, then follow for additional time
directive sample -2*24*3600, 1*24*3600.0 1000
directive simulation deterministicstiff
directive plot Me-P1; Me-P2  //specify output data written to file
// Declare all of the parameters that are to be inferred
directive parameters [ in1 = 0.0
                     ; in2 = 0.0; upreg1 =0.0; upreg2 = 0.0
                     ; upfactor1 ,(1.0 ,10000.0) ,1.0 ,log ,random
                     ; upfactor2 ,(1.0 ,10000.0) ,1.0 ,log ,random
                     ; offset1 ,(0.0 ,5000.0) ,1.0 ,real ,random
                     ; offset2 ,(0.0 ,5000.0) ,1.0 ,real ,random
                     ; offset3 ,(0.0 ,5000.0) ,1.0 ,real ,random
                     ; offset4 ,(0.0 ,15000.0) ,1.0 ,real ,random
                     ; offset5 ,(0.0 ,5000.0) ,1.0 ,real ,random
                     ; offset6 ,(0.0 ,5000.0) ,1.0 ,real ,random
                     ; offset7 ,(0.0 ,5000.0) ,1.0 ,real ,random
                     ; offset8 ,(0.0 ,15000.0) ,1.0 ,real ,random
                     ; s1 ,(1e-3 ,10000.0) ,1.0 ,log ,random
                     ; s2 ,(1e-3 ,10000.0) ,1.0 ,log ,random
                     ; sf1 ,(1e-4 ,10000.0) ,1.0 ,log ,random
                     ; sf2 ,(1e-4 ,10000.0) ,1.0 ,log ,random
                     ; sf3 ,(1e-4 ,10000.0) ,1.0 ,log ,random
                     ; b2, (1e-12,1e-6) ,2.755e-10,log ,random
                     ; g0, (0.0, 100000.0) ,25000.0 ,log ,random
```

```
                                ]
//Auto-generated sweep 151209
directive sweep mysweep1 = {upreg1 = [0.0], upreg2 = [0.0],
(in1,in2) = [(78.00,88.00),(171.00,134.00),(519.00,153.00),
(1809.00,141.00),(5641.00,138.00),(17934.00,135.00),
(58085.00,137.00),(139657.00,142.00),(78.00,609.00),
(181.00,588.00),(578.00,570.00),(1808.00,612.00),
(5475.00,625.00),(17452.00,632.00),(55916.00,610.00),
(139077.00,615.00),(79.00,1842.00),(178.00,1841.00),
(588.00,1800.00),(1861.00,1855.00),(5619.00,1957.00),
(17284.00,1991.00),(55487.00,1961.00),(137132.00,1986.00),
(81.00,5665.00),(177.00,5697.00),(580.00,5479.00),
(1935.00,5567.00),(5929.00,5863.00),(17868.00,6294.00),
(55739.00,6583.00),(140036.00,6509.00),(79.00,18045.00),
(175.00,17527.00),(570.00,17486.00),(1966.00,17219.00),
(6353.00,17338.00),(19056.00,18146.00),(56173.00,18955.00),
(141468.00,19820.00),(74.00,57977.00),(173.00,54937.00),
(542.00,54758.00),(1965.00,52819.00),(6431.00,51047.00),
(19215.00,52628.00),(58684.00,54185.00),(142756.00,55668.00),
(65.00,141871.00),(174.00,137576.00),(573.00,135604.00),
(1740.00,141735.00),(6424.00,134635.00),(19446.00,134065.00),
(60533.00,134468.00),(147211.00,136278.00)] }
//Auto-generated sweep 151209
directive sweep mysweep2 = {upreg1 = [1.0], upreg2 = [0.0],
(in1,in2) = [(81.00,104.00),(167.00,153.00),(525.00,171.00),
(1760.00,156.00),(5465.00,147.00),(17538.00,142.00),
(52483.00,141.00),(137115.00,154.00),(80.00,561.00),
(179.00,541.00),(573.00,551.00),(1763.00,628.00),
(5072.00,650.00),(15504.00,611.00),(56050.00,621.00),
(122795.00,672.00),(79.00,1813.00),(172.00,1738.00),
(639.00,1727.00),(1851.00,1770.00),(5429.00,1934.00),
(16911.00,1992.00),(48800.00,2013.00),(136065.00,2106.00),
(81.00,5521.00),(169.00,5465.00),(649.00,5082.00),
(1966.00,5162.00),(5960.00,5600.00),(17125.00,6176.00),
(49736.00,6579.00),(135853.00,6028.00),(78.00,16649.00),
(165.00,16226.00),(634.00,16788.00),(2074.00,15752.00),
```

```
(6291.00,16024.00),(18332.00,17139.00),(52784.00,18437.00),
(129274.00,19557.00),(72.00,54056.00),(160.00,51805.00),
(553.00,47510.00),(1890.00,48063.00),(6433.00,47686.00),
(20231.00,48412.00),(55721.00,49739.00),(133010.00,54457.00),
(67.00,138580.00),(147.00,132303.00),(512.00,147005.00),
(1921.00,149353.00),(5965.00,127421.00),(20165.00,124047.00),
(61527.00,131767.00),(140315.00,139052.00)] }
//Auto-generated sweep 151209
directive sweep mysweep3 = {upreg1 = [0.0], upreg2 = [0.0],
(in1,in2) = [(78.00,88.00),(172.00,133.00),(518.00,155.00),
(1802.00,143.00),(5624.00,139.00),(18056.00,136.00),
(57925.00,140.00),(141163.00,143.00),(78.00,615.00),
(183.00,594.00),(568.00,570.00),(1800.00,617.00),
(5438.00,605.00),(16929.00,608.00),(57058.00,596.00),
(137510.00,579.00),(78.00,1852.00),(179.00,1826.00),
(590.00,1814.00),(1884.00,1859.00),(5623.00,1944.00),
(17312.00,2001.00),(53179.00,1996.00),(139440.00,1941.00),
(81.00,5742.00),(179.00,5731.00),(570.00,5643.00),
(1925.00,5517.00),(5979.00,5819.00),(17692.00,6222.00),
(55236.00,6610.00),(140745.00,6527.00),(80.00,17870.00),
(175.00,17617.00),(553.00,17492.00),(1981.00,16890.00),
(6371.00,17624.00),(18898.00,18577.00),(57001.00,18991.00),
(141627.00,19390.00),(76.00,58367.00),(172.00,55549.00),
(552.00,54875.00),(1999.00,53632.00),(6391.00,51420.00),
(19547.00,51939.00),(59146.00,53311.00),(141015.00,56074.00),
(69.00,139994.00),(170.00,136697.00),(539.00,140727.00),
(1782.00,133716.00),(6000.00,134770.00),(20105.00,132830.00),
(59194.00,133914.00),(144102.00,139359.00)] }
//Auto-generated sweep 151209
directive sweep mysweep4 = {upreg1 = [1.0], upreg2 = [0.0],
(in1,in2) = [(81.00,101.00),(167.00,148.00),(526.00,170.00),
(1766.00,153.00),(5518.00,142.00),(17506.00,133.00),
(54797.00,133.00),(128159.00,144.00),(79.00,569.00),
(180.00,542.00),(589.00,554.00),(1743.00,623.00),
(5019.00,649.00),(16056.00,671.00),(53406.00,576.00),
(111982.00,576.00),(78.00,1801.00),(173.00,1736.00),
```

```
98 (641.00,1688.00),(1867.00,1788.00),(5327.00,1892.00),
99 (16330.00,2048.00),(46630.00,2085.00),(136859.00,1752.00),
100 (83.00,5437.00),(168.00,5453.00),(623.00,5100.00),
101 (1961.00,5192.00),(5860.00,5510.00),(16914.00,5815.00),
102 (50490.00,6378.00),(136732.00,6150.00),(80.00,16694.00),
103 (163.00,16634.00),(594.00,17412.00),(2018.00,15698.00),
104 (6342.00,16022.00),(17922.00,16721.00),(53210.00,18537.00),
105 (123770.00,20402.00),(75.00,54042.00),(171.00,51174.00),
106 (578.00,53977.00),(1948.00,45585.00),(6633.00,43022.00),
107 (18711.00,49379.00),(53532.00,49956.00),(131133.00,54806.00),
108 (71.00,132070.00),(177.00,135831.00),(539.00,133865.00),
109 (1826.00,135788.00),(7907.00,172254.00),(19643.00,131200.00),
110 (55280.00,143953.00),(145791.00,142991.00)] }
111 //Auto-generated sweep 161102
112 directive sweep mysweep5 = {upreg2 = [0.0], upreg1 = [0.0],
113 (in1,in2) = [(75.00,81.00),(156.00,121.00),(534.00,137.00),
114 (1784.00,125.00),(5625.00,119.00),(18014.00,112.00),
115 (57576.00,112.00),(140549.00,113.00),(76.00,593.00),
116 (171.00,585.00),(582.00,587.00),(1744.00,645.00),
117 (5283.00,637.00),(17805.00,629.00),(54386.00,592.00),
118 (140822.00,607.00),(76.00,1841.00),(162.00,1832.00),
119 (634.00,1742.00),(1846.00,1870.00),(5534.00,1962.00),
120 (17434.00,2030.00),(54600.00,1999.00),(145170.00,1942.00),
121 (76.00,5670.00),(162.00,5627.00),(632.00,5414.00),
122 (1954.00,5452.00),(5990.00,5903.00),(17701.00,6238.00),
123 (54439.00,6415.00),(140807.00,6665.00),(78.00,17753.00),
124 (165.00,17574.00),(586.00,17374.00),(1935.00,17288.00),
125 (6405.00,17320.00),(18494.00,18166.00),(55530.00,18936.00),
126 (140319.00,19219.00),(76.00,57824.00),(164.00,55975.00),
127 (572.00,54368.00),(1946.00,54511.00),(6490.00,50836.00),
128 (19431.00,52494.00),(57441.00,53148.00),(142745.00,55356.00),
129 (77.00,137396.00),(168.00,142074.00),(566.00,148459.00),
130 (1838.00,139579.00),(6080.00,142822.00),(19399.00,136127.00),
131 (60762.00,137339.00),(147539.00,140172.00)] }
132 //Auto-generated sweep 161102
133 directive sweep mysweep6 = {upreg2 = [1.0], upreg1 = [0.0],
```

```
(in1,in2) = [(73.00,85.00),(146.00,122.00),(550.00,130.00),
(1750.00,111.00),(5453.00,92.00),(18025.00,66.00),
(52313.00,55.00),(128732.00,41.00),(74.00,579.00),
(165.00,561.00),(599.00,584.00),(1709.00,641.00),
(5023.00,676.00),(16251.00,664.00),(51272.00,568.00),
(125013.00,644.00),(73.00,1803.00),(156.00,1766.00),
(642.00,1734.00),(1863.00,1827.00),(5317.00,1933.00),
(16612.00,2041.00),(48804.00,2066.00),(131693.00,1894.00),
(72.00,5573.00),(151.00,5433.00),(654.00,5180.00),
(1959.00,5405.00),(5762.00,5730.00),(17010.00,6229.00),
(48932.00,6745.00),(138189.00,6372.00),(75.00,17585.00),
(148.00,17276.00),(667.00,16754.00),(1954.00,16027.00),
(6175.00,16282.00),(18134.00,17341.00),(49905.00,18898.00),
(128404.00,17083.00),(70.00,52460.00),(145.00,49252.00),
(632.00,46090.00),(1997.00,45525.00),(6625.00,47476.00),
(18762.00,47137.00),(53685.00,49115.00),(127353.00,52637.00),
(80.00,135953.00),(137.00,136504.00),(524.00,130670.00),
(2084.00,133443.00),(6290.00,121134.00),(15737.00,138674.00),
(53091.00,120365.00),(133581.00,127439.00)] }
//Auto-generated sweep 161102
directive sweep mysweep7 = {upreg2 = [0.0], upreg1 = [0.0],
(in1,in2) = [(74.00,81.00),(155.00,120.00),(536.00,137.00),
(1783.00,124.00),(5672.00,120.00),(18128.00,117.00),
(57691.00,119.00),(140604.00,123.00),(75.00,591.00),
(171.00,586.00),(586.00,583.00),(1758.00,640.00),
(5363.00,638.00),(17437.00,619.00),(58561.00,627.00),
(141890.00,575.00),(75.00,1855.00),(164.00,1813.00),
(630.00,1756.00),(1865.00,1871.00),(5542.00,1978.00),
(17695.00,1988.00),(54462.00,1996.00),(136606.00,1931.00),
(73.00,5696.00),(161.00,5581.00),(627.00,5431.00),
(1943.00,5415.00),(5893.00,5883.00),(17582.00,6265.00),
(54693.00,6468.00),(144240.00,6583.00),(76.00,18018.00),
(161.00,17654.00),(600.00,17617.00),(2026.00,17196.00),
(6264.00,17302.00),(18709.00,18076.00),(55521.00,18904.00),
(142386.00,19278.00),(72.00,56921.00),(165.00,55834.00),
(565.00,56076.00),(1942.00,51130.00),(6568.00,51921.00),
```

```
(19657.00,53419.00),(58130.00,54120.00),(142698.00,56280.00),
(74.00,137189.00),(162.00,139084.00),(533.00,141118.00),
(1961.00,140302.00),(6503.00,138611.00),(19894.00,134501.00),
(62051.00,136911.00),(145938.00,139125.00)] }
//Auto-generated sweep 161102
directive sweep mysweep8 = {upreg2 = [1.0], upreg1 = [0.0],
(in1,in2) = [(71.00,83.00),(144.00,118.00),(560.00,128.00),
(1746.00,117.00),(5445.00,105.00),(17848.00,96.00),
(51283.00,90.00),(134623.00,95.00),(70.00,577.00),
(166.00,555.00),(602.00,570.00),(1747.00,630.00),
(5135.00,667.00),(17040.00,666.00),(49637.00,683.00),
(109247.00,371.00),(65.00,1796.00),(160.00,1721.00),
(639.00,1770.00),(1881.00,1803.00),(5349.00,1944.00),
(16302.00,2033.00),(48977.00,1975.00),(144868.00,2112.00),
(58.00,5673.00),(157.00,5348.00),(651.00,5275.00),
(1947.00,5360.00),(5858.00,5651.00),(17048.00,6057.00),
(48248.00,6388.00),(138192.00,6099.00),(50.00,17215.00),
(156.00,16420.00),(658.00,16032.00),(2005.00,16184.00),
(6242.00,16517.00),(18644.00,17876.00),(49785.00,18293.00),
(128876.00,18794.00),(45.00,49791.00),(140.00,50981.00),
(673.00,49821.00),(1948.00,48824.00),(6340.00,47012.00),
(18955.00,50527.00),(54026.00,50145.00),(127979.00,52688.00),
(40.00,130908.00),(155.00,126501.00),(742.00,148286.00),
(1805.00,145284.00),(6918.00,139213.00),(20599.00,122126.00),
(51805.00,141182.00),(148245.00,136779.00)] }
// We add a scaled quantity of each input. The scale factors are s1
   and s2 for 151209 and 161102 respectively. Event occurs at t=0
directive event I1 s1*in1 @ 0.0
directive event I2 s2*in2 @ 0.0
// Specify the relationship between model simulation and data, and
   MCMC settings
directive fit { mysweep1; SSL_ASN_151209_merged; [Me-P1*SF3+OFF1] }
directive fit { mysweep2; SSL_ASN_151209_IFN_merged; [Me-P1*SF3+OFF2] }
directive fit { mysweep3; ASN_151209_merged_new2; [Me-P2*SF2+OFF3] }
directive fit { mysweep4; ASN_151209_IFN_merged_new2; [Me-P2*SF2+OFF4]
   }
```

```
203 directive fit { mysweep5; SSL_ASN_161102_merged; [Me~P1*SF1+OFF5] }
204 directive fit { mysweep6; SSL_ASN_161102_IFN_merged; [Me~P1*SF1+OFF6] }
205 directive fit { mysweep7; ASN_161102_merged_new; [Me~P2*SF2+OFF7] }
206 directive fit { mysweep8; ASN_161102_IFN_merged2; [Me~P2*SF2+OFF8] }
207 directive fit_run { burnin = 20000; samples = 50000; thin = 10;
        noisemodel = 1 }
208 // Standard parameter values
209 rate dMe = 5.193e-5;
210 rate b0 = 2.755e-10; //binding rate of self peptide
211 rate b1 = 2.755e-10; //binding rate of SSL
212 rate c = 8.302928e-8;
213 rate dP = 0.13;
214 rate uT = 1.184643e-6;
215 rate vT = 0.0011091974705091;
216 rate bT = 1.662768e-9;
217 rate gT = 1505;
218 rate dT = 0.001725968;
219 rate e = 7.071e-4;//0.1141804;
220 rate gM = 150.5;
221 rate dM = 7.9892e-5;
222 rate q = 21035;
223 // Specialising peptide supply and off-rates
224 rate u1 = 2.8E-05; //SSL off-rate
225 rate u2 = 5.2e-05; //ASN off-rate
226 rate u0 = 1e-4; // self off-rate
227
228 spec P1; spec I1;
229 spec P2; spec I2;
230 spec P0; spec I0;
231
232 // You have to use a species in the fit directive, so just initialise
        a constant species with a parameterized value.
233 // SF1 = scale factor number of copies of SSL to MFI for 161102,
234 // SF3 = scale factor number of copies of SSL to MFI for 151209
235 // SF2 = scale factor number of copies of ASN to MFI for both 161102
        and 151209
```

```
236 init SF1 sf1 |
237 init SF2 sf2 |
238 init SF3 sf3 |
239 init OFF1 offset1 |
240 init OFF2 offset2 |
241 init OFF3 offset3 |
242 init OFF4 offset4 |
243 init OFF5 offset5 |
244 init OFF6 offset6 |
245 init OFF7 offset7 |
246 init OFF8 offset8 |
247 // Standard modular definition of the MHC model module Pep(spec Pi,Ii;
      rate gi,ui,bi){
248   Ii  ~->{gi} Pi | Pi ->{dP} |
249   M + Pi <->{bi}{ui} M-Pi |
250   T-M + Pi <->{c}{ui*q} T-M-Pi |
251   T-M-Pi ->{vT} T + M-Pi |
252   M-Pi ->{e} Me-Pi |
253   Me-Pi ->{ui} Me
254 };
255 <->[gM + upreg1*upfactor1 + upreg2*upfactor2]{dM} M |
256 <->[10*(gM + upreg1*upfactor1 + upreg2*upfactor2)]{dT} T |
257 T + M <->{bT}{uT} T-M |
258 Me ->{dMe} |
259
260 // Instantiate the three peptides (target, competitor, background)
      with BG supplied at rate g0
261 Pep(P1,I1,1.0,u1,b1) |
262 Pep(P2,I2,1.0,u2,b2) |
263 init I0 1 | Pep(P0,I0,g0,u0,b0)
```

Appendix B

HeLa Cell Project: Matlab Code

```
%% Script to load in peptide files and simulate deterministically

clear
close all
alleleStr = {'A6802','B4403','B2705','B5801'}; %List of alleles to
    choose from
allele = 1; % if = 1 then allele = B4402, if = 2 then allele = B2705,
    if = 3 then allel =B4405
disp('Loading sequences and copy numbers')
tic
load('Total_mut.mat')
%res = T_mut.res;
%load('Total.mat');
%upi = T.UPI_new;
ipi = totaldata.IPIA;
proteins = totaldata.alldataseq;
half_lives = totaldata.hlMatch;
copies = totaldata.Intensity;
res = totaldata.residues;
filepath = 'path\to\HUMAN_9606_idmapping_selected.tab\';
delimiter = ',';startRow = 3;formatSpec =
    '%s%f%f%f%f%s%f%f%f%f%f%f%[^\n\r]';
N=length(ipi);
peps = struct;
for i = find(matchindex(:,1)>0)
```

```
filename = [filepath ipi{i} 'mut_new.dat'];
fileID = fopen(filename,'r');
dataArray = textscan(fileID, formatSpec, 'Delimiter', delimiter,
    'HeaderLines' ,startRow-1, 'ReturnOnError', false);
fclose(fileID);
pstart = dataArray{:, 3};
pend = dataArray{:, 4};
peptide = dataArray{:, 6};
protcleave = dataArray{:, 7};
ic50 = dataArray{:, 12};
peps(i).peptide = peptide;
peps(i).protcleave = protcleave;
peps(i).ic50 = ic50;
peps(i).pstart = pstart;
peps(i).pend = pend;
end
toc
%% Remove low affinity peptides
disp('Preparing off-rates and supply rates')
% Count the number of peptides
np=0;
for i = 1:numel(peps)
  %for j = 1:length(res{i})
     np = np + length(peps(i).peptide);
  %end
end
n = length(copies);
bP = 10^3;
% Assign off-rates and peptide sequences
p = 0;
seq = cell(np,1);
data = zeros(np,6);
protID = cell(np,1);
residue = zeros(np,4);
for i = 1: numel(peps)
     nu = length(peps(i).ic50);
```

```
58      data(p+1:p+nu,1) = peps(i).ic50*bP/10^9;
59      data(p+1:p+nu,2) = copies(i);
60      data(p+1:p+nu,3) = half_lives(i);
61      data(p+1:p+nu,4) = peps(i).protcleave;
62      data(p+1:p+nu,5) = peps(i).pstart;
63      data(p+1:p+nu,6) = peps(i).pend;
64      seq(p+1:p+nu) = peps(i).peptide;
65      protID(p+1:p+nu) = ipi(i);
66      residue(p+1:p+nu,1)=res(i,1);
67      residue(p+1:p+nu,2)=res(i,2);
68      residue(p+1:p+nu,3)=res(i,3);
69      residue(p+1:p+nu,4)=res(i,4);
70      p = p + nu;
71  end
72
73  %% Trim
74  disp('Trimming peptides with large off-rates')
75  uMax = 1e-2;
76  locs = find(data(:,1)<=uMax);
77  data = data(locs,:);
78  seq = seq(locs);
79  n0 = length(data(:,1));
80  residue = residue(locs,:);
81  %% Prepare parameters
82  load bOpt  % load parameters p and pvars
83  p1.b = 2.755e-10;% p1.bs(allele); %Assign peptide binding rate
        according to allele
84  nTAP = 10000; %Number of TAP molecules per cell
85  p0 = p1; p0.gT = 0;
86  Ptot = 2*nTAP/p1.dP; % Total peptide
87
88  %% Simulate the full system
89  disp('Simulating ODEs')
90  xInit0 = zeros(4+3*n0,1);
91  odeopts = odeset('Jacobian',@JacobianFixP_newND);
92  tic
```

```
93 [t0,x0] = ode15s(@PLodesFixP_newND,[0
       10*24*3600],xInit0,odeopts,p1,data,n0,Ptot);
94 toc
95 MeP0 = x0(end,4+2*n0:3+3*n0);
96 %Sort data and save
97 [sort_MeP0, in] = sort(MeP0, 'descend');
98 sort_protID = protID(in);
99 sort_seq = seq(in);
100 sort_copies = data(in,2);
101 sort_offrates = data(in,1);
102 sort_deg = data(in,3);
103 sort_prot = data(in,4);
104 sort_start = data(in,5);
105 sort_end = data(in,6);
106 sort_residue = residue(in,:);
107 output_sort = table(sort_MeP0', sort_seq, sort_protID, sort_copies,
       sort_offrates, sort_deg, sort_prot, sort_start,
       sort_end,sort_residue);
108 writetable(output_sort,[alleleStr{allele} '_mut_sort_new.txt']);
109
110 return
```

Appendix C

HIV Project: Matlab Code

```
function
    [t,x,nSelf,n,genes,Sequence]=HIVAll_cvodes_old(allele)%,alleles)

%READ IN TOP 1% DATA FOR EACH ALLELE
name = ['top' allele];
filename = ['path\to\allele\file' allele '.xlsx'];
[data, txt, raw]= xlsread(filename, 'Sheet1');
raw(cellfun(@(x) ~isempty(x) && isnumeric(x) && isnan(x),raw)) = {''};
cellVectors = raw(:,[1,5,6]);
%ASSIGN DATA FROM FILE
Allele = cellVectors(2:end,1);
Start = data(:,1);
End = data(:,2);
Length = data(:,3);
Sequence = cellVectors(2:end,2);
genes = cellVectors(2:end,3);
ProteasomeScore = data(:,6);
TAPScore = data(:,7);
MHCScore = data(:,8);
ProcessingScore = data(:,9);
TotalScore = data(:,10);
MHCIC50nM = data(:,11);
Abundance = data(:,12);
%% Create output variable
%% Clear temporary variables
```

```
clearvars data raw cellVectors;
%CALCULATE PARAMETER VECTORS
proteins = {'REV','TAT','NEF','GAG','POL','ENV','VIF','VPR','VPU'};
u = MHCIC50nM*1e3/10^9;
u=u*3600;
n=length(u);
g=zeros(n,1);
g(:)=0.08*3600;
bP=3600*2.8871e-9;
bPeps=zeros(n,1);bPeps(:)=bP;
Abundance = 10.^ProteasomeScore;
ps = Abundance/1000;
pepfilename = [allele 'dPc.xlsx'];
[pepdata, ~, pepraw] = xlsread(pepfilename);
pepraw(cellfun(@(x) ~isempty(x) && isnumeric(x) && isnan(x),pepraw)) =
    {''};
pepcellVectors = pepraw(:,1);
%Calculate peptide degradation
Peptide1 = pepcellVectors(2:end,1);
p5 = pepdata(:,1);
p30 = pepdata(:,2);
p2 = pepdata(:,3);
clearvars pepdata pepraw pepcellVectors;

dPeps = peptide_deg(p5,p30,p2);
dPeps=dPeps(:)*3600;
%ASSIGN SELF PEPTIDE PARAMETERS
ull=60*60*1e-2;
ul=60*60*1e-3;
um=60*60*1e-4;
uh=60*60*1e-5;
uSelf=[ull;ul;um;uh];
G=20000;

gll=60*60*G*0.5/100;
gl=60*60*G*98.5/100;
```

```
gm=60*60*G*0.5/100;
gh=60*60*G*0.5/100;
gSelf=[gll;gl;gm;gh];
pSelf=[0.03;0.03;0.03;0.03];
nSelf=length(uSelf);

ns9=4+(5*nSelf)+9;
TP=6+(12*nSelf);
TP2=4+(5*nSelf);
nss=TP+147+n;
NN=nss+n;
pv=0;
% ASSIGN INITIAL CONCENTRATIONS
% HOWEVER IF PV=0 THEN THESE ARE ALL SET TO 0
x0=zeros(1,TP2+49+(5*n));
x0(1,TP2+35)=2000*pv; %Gag
x0(1,TP2+36)=(2000/20)*pv; %GagPol
x0(1,TP2+37)=33*pv;%282;%Env
x0(1,TP2+38)=101*pv;%Vif
x0(1,TP2+39)=285*pv; %Vpr
x0(1,TP2+40)=2000*0.01*pv;%Vpu
x0(1,TP2+41)=150*pv;%2000*0.01;%Nef
x0(1,TP2+5)=0;%2000*0.01;%Rev
x0(1,TP2+6)=0;%2000*0.01;%Tat
%DEFINE ODE OPTIONS TO INCLUDE JACOBIAN
options=odeset('Jacobian',@jacobian_Cvodes);
% DEFINE START AND END TIMES OF SIMULATON
tstart=0;
tfinal=72;
%SOLVE SYSTEM OF ODES
[t,x]=ode15s(@HIV1_ALLeff_old, [tstart,
    tfinal],x0,options,u,g,ps,genes,uSelf,pSelf,gSelf,nSelf,dPeps,bPeps);
sim = table(t,x);
%filename = [allele '_sim'];
%writetable(sim, filename);
end
```

```
95 %-------------------------------------------------------------------------
96 function dxdt =
       HIV1_ALLeff_old(t,x,u,g,ps,genes,uSelf,pSelf,gSelf,nSelf,dPeps,bPeps)
97
98 [uT, bT, gT, dT, q, c, uTv, e, gM, dM, dMe, dPc,dP, bP]=MHC_parameters;
99 %Parameters%
100
101 pv=1;%number of proviruses
102 tb=15; %basal rate of transcription per hour
103 tadd=1485; %increase in viral RNA transcription by Tat transactivation
       per hour
104
105 aTat=28.57;%Association constant of Tat with TAT (/muM)
106 ka1=0.0132; %high affinity Rev binding (association constant)
107 kd1=60*60*(3.0e-5); %high affinity Rev binding dissociation constant
       (h)^-1
108 kai=0.0233; %low affinity Rev binding association constant
109 kdi=60*60*(3.8e-2); %low affinity Rev binding dissociation constant
       (h)^-1
110 sn=12; %maximum number of Rev per RRE
111 th=7; %threshold level of Rev for mRNA nuclear export
112 sp=2.5; %splicing rate constant
113 kexp=2.082; %export rate constant
114 kimp=20.8200; %import rate constant
115 kdeg=0.1740; %rate constant of RNA degradation in nucleus and cytoplasm
116 tr=270; %steady-state translation rate
117 kRevn=0.0432; %degradation rate of Rev in nucleus
118 kRevc=0.0174; %degradation rate of Rev in cytoplasm
119 kTatc=0.0174; %degradation rate of Tat in cytoplasm
120 fRNAr=0.19; %fraction of rev mRNA in fullyspliced mRNA
121 fRNAt=0.05; %fraction of tat mRNA in singly and fully spliced mRNAs
122 fRev=0.5; %probability for rev mRNA to encode Rev
123 fTat=1; %probability for tat mRNA to encode Tat
124 d=0.8; %splicing delay factor
125 factor=0.000014687; %conversion from muM^-1 to /molecule
126 factor2=1; %conversion from M^-1 to /molecule
```

```
127 kbud=0.08; %budding rate
128 fGag=0.95; %fraction of Gag translated from FC
129 fPol=0.05; %fraction of Pol translated from FC
130 nGagvirion=2000;  %number of Gag in virion (this is actually for 4900
        but was set as this in the Wang and LuHa model)
131 kTat=0.1733; %Rate of degradation of Tat per hour
132 kGag=0.1054; %Rate of degradation of Gag per hour
133 kVif=0.4673; %Rate of degradation of Vif per hour
134 kEnv=0.02; %Rate of degradation of Env per hour
135 kVpr=0.0346; %Rate of degradation of Vpr per hour
136 kVpu=0.08664; %Rate of degradation of Vpu per hour
137 aGagVif=2e-6; %Association constant of Gag with Vif (/molecule)
138 n=length(u); %number of peptides
139 %GET PROTEINS FROM WHICH PEPTIDES COME FROM
140 [nGag,nVpu,nVpr,nVif,nRev,nTat,nGagPol,nEnv,nNef]=length_genes(genes);
141
142 ns9=4+(5*nSelf)+9;
143 TP=6+(12*nSelf);
144 TP2=4+(5*nSelf);
145 nss=TP+147+n;
146 NN=nss+n;
147
148 r=zeros(NN+(10*n),1);
149 %Production/degradation of M and T
150
151 r(1)= gM; % supply of M
152 r(2)= dM*x(1); % degradation of M where x(1)=M
153 r(3)= gT; % supply of T
154 r(4) =dT*x(2); % degradation of T
155 r(5)= bT*x(1)*x(2); % binding of T to M where x(2) is T
156 r(6)= uT*x(3); % unbinding of T from M where x(3)=M-T
157 r(7:6+nSelf)=0;%%pSelf(:)*prt*SelfProt; %production of self peptides
        in the cytoplasm (may need to be tweeked)
158 r(6+nSelf+1:6+(2*nSelf))=0;%%dPc*x(5:4+nSelf); %x(1:nSelf)=Self
        Peptides in cytoplasm
159 r(6+(2*nSelf)+1:6+(3*nSelf))=gSelf(:);%.*x(5:4+nSelf); %
```

```
    transportation of self peptides to the ER from the cytoplasm
r(6+(3*nSelf)+1:6+(4*nSelf))=dPc(:).*x(4+nSelf+1:4+(2*nSelf)); %
    degradation of self peptides in the ER,
    x(4+nSelf+1:4+2*nSelf)=self peptides in the ER
r(6+(4*nSelf)+1:6+(5*nSelf))=bP*x(4+nSelf+1:4+(2*nSelf))*x(1);
    %binding of self peptides to M, x(1)=M
r(6+(5*nSelf)+1:6+(6*nSelf))=uSelf(:).*x(4+(2*nSelf)+1:4+(3*nSelf));
    %unbinding where x(4+(2*nSelf)+1:4+3*nSelf)=M-SelfPep
r(6+(6*nSelf)+1:6+(7*nSelf))=c*x(4+nSelf+1:4+(2*nSelf))*x(3); %self
    pep binding to MT, where x(3)=M-T FOR NOW THIS WILL CHANGE
r(6+(7*nSelf)+1:6+(8*nSelf))=q*uSelf(:).*x(4+(3*nSelf)+1:4+(4*nSelf));
    % unbinding of selfPep from M-T, where
    x(4+(3*nSelf+1:4+4*nSelf))=M-T-SelfPep
r(6+(8*nSelf)+1:6+(9*nSelf))=uTv*x(4+(3*nSelf)+1:4+(4*nSelf)); %
    unbinding of T from M-T-SelfPep
r(6+(9*nSelf)+1:6+(10*nSelf))=e*x(4+(2*nSelf)+1:4+(3*nSelf));
    %egression of M-SelfPep where x((2*nSelf)+1:3*nSelf)=M-SelfPep
r(6+(10*nSelf)+1:6+(11*nSelf))=uSelf(:).*x(4+(4*nSelf)+1:4+(5*nSelf));
    %unbinding of SelfPep on cell surface
    x((4*nSelf)+1:5*nSelf)=Me-SelfPep
r(6+(11*nSelf)+1:6+(12*nSelf))=0;% dMe*x(4); % where x(4) = Me

%Viral Proteins- their intial levels will be set
%SP=6+(12*nSelf);
TP=6+(12*nSelf);%SPN+(11*n);
TP2=4+(5*nSelf);
r(TP+1)=tb +(tadd*aTat*factor*x(TP2+8)/(1+aTat*factor*x(TP2+8)))*pv;
    %transcription of FN
r(TP+2)=sp*x(TP2+1); %splicing of FN
r(TP+3)=kdeg*x(TP2+1); %degradation of FN
r(TP+4)=sp*x(TP2+2); %splicing of SN
r(TP+5)=kdeg*x(TP2+2); %degradation of SN
r(TP+6)=kexp*x(TP2+3); %export of MN
r(TP+7)=kdeg*x(TP2+3); %degradation of MN
r(TP+8)=kdeg*x(TP2+4); %degradation of MC
r(TP+9)=fRNAr*tr*fRev*x(TP2+4); %translation of Rev protein
```

```
183 r(TP+10)=kRevc*x(TP2+5); %degradation of Rev in cytoplasm
184 r(TP+11)=kimp*x(TP2+5); %import of Rev to nucleus
185 r(TP+12)=fRNAt*tr*fTat*x(TP2+4); %translation of Tat protein
186 r(TP+13)=kRevc*x(TP2+6); %degradation of Tat in cytoplasm
187 r(TP+14)=kimp*x(TP2+6); %import of Tat to nucleus
188 r(TP+15)=kRevn*x(TP2+7); %degradation of Revn
189 r(TP+16)=kRevn*x(TP2+8); %degradation of Tatn
190 r(TP+17)=kexp*x(TP2+7); %export of Revn
191 r(TP+18)=ka1*factor2*x(TP2+1)*x(TP2+7); %binding of Rev to FN
192 r(TP+19)=kd1*x(TP2+9); %unbinding of Rev from FNR1
193 r(TP+20)=(1-d)*sp*x(TP2+9); %splicing of FNR1
194 r(TP+21)=kdeg*x(TP2+9); %degradation of FNR1 and release of 1 Rev
195 r(TP+22)=(1-d)*sp*x(TP2+10); %splicing of SNR1 to produced MN and
    release 1 Rev
196 r(TP+23)=kdeg*x(TP2+10); %degradation of SNR1 to release 1 Rev
197 r(TP+24)=ka1*factor2*x(TP2+2)*x(TP2+7); %binding of Rev to SN
198 r(TP+25)=kd1*x(TP2+10); %unbinding of Rev from SNR1
199 r(TP+26)=kai*factor2*x(TP2+9)*x(TP2+7); % binding of Rev to FNR1 to
    produce FNR2
200 r(TP+27)=kdi*x(TP2+11); %unbinding of Rev from FNR2
201 r(TP+28)=(1-d)*sp*x(TP2+11); %splicing of FNR2 to produce SNR2
202 r(TP+29)=kdeg*x(TP2+11); %degradation of FNR2 to release 2 Revs
203 r(TP+30)=(1-d)*sp*x(TP2+12); %splicing of SNR2 to produce MN and 2*Revs
204 r(TP+31)=kdeg*x(TP2+12); %degradation of SNR2 to release 2 Revs
205 r(TP+32)=kai*factor2*x(TP2+10)*x(TP2+7); %binding of Rev to SNR1 to
    produce SNR2
206 r(TP+33)=kdi*x(TP2+12); %unbinding of 1 Rev from SNR2
207 r(TP+34)=kai*factor2*x(TP2+11)*x(TP2+7); %binding of Rev to FNR2 to
    produce FNR3
208 r(TP+35)=kdi*x(TP2+13); %unbinding of Rev from FNR3
209 r(TP+36)=(1-d)*sp*x(TP2+13); %splicing of FNR3 to produce SNR3
210 r(TP+37)=kdeg*x(TP2+13); %degradation of SNR3 to release 3 Revs
211 r(TP+38)=(1-d)*sp*x(TP2+14); %splicing of SNR3 to produce MN and 3*Revs
212 r(TP+39)=kdeg*x(TP2+14); %degradation of SNR3 to release 3 Revs
213 r(TP+40)=kai*factor2*x(TP2+12)*x(TP2+7); %binding of Rev to SNR2 to
    produce SNR3
```

```
214 r(TP+41)=kdi*x(TP2+14); %unbinding of 1 Rev from SNR3
215 r(TP+42)=kai*factor2*x(TP2+13)*x(TP2+7); %binding of Rev to FNR3 to
      produce FNR4
216 r(TP+43)=kdi*x(TP2+15); %unbinding of Rev from FNR4
217 r(TP+44)=(1-d)*sp*x(TP2+15); %splicing of FNR4 to produce SNR4
218 r(TP+45)=kdeg*x(TP2+15); %degradation of FNR4 to release 4 Revs
219 r(TP+46)=(1-d)*sp*x(TP2+16); %splicing of SNR4 to produce MN and 4 Revs
220 r(TP+47)=kdeg*x(TP2+16); %degradation of SNR4 to produce 4 Revs
221 r(TP+48)=kai*factor2*x(TP2+14)*x(TP2+7); %binding of Rev to SNR3 to
      produce SNR4
222 r(TP+49)=kdi*x(TP2+16); %unbinding of Rev from SNR4 to produce SNR3
223 r(TP+50)=kai*factor2*x(TP2+15)*x(TP2+7); %binding of Rev to FNR4 to
      produce FNR5
224 r(TP+51)=kdi*x(TP2+17); %unbinding of Rev from FNR5 to produce FNR4
225 r(TP+52)=(1-d)*sp*x(TP2+17); %splicing of FNR5 to produce SNR5
226 r(TP+53)=kdeg*x(TP2+17); %degradation of FNR5 to release 5 Revs
227 r(TP+54)=(1-d)*sp*x(TP2+18); %splicing of SNR5 to produce MN and 5 Revs
228 r(TP+55)=kdeg*x(TP2+18); %degradation of SNR5 to produce 5 Revs
229 r(TP+56)=kai*factor2*x(TP2+16)*x(TP2+7); %binding of Rev to SNR4 to
      produce SNR5
230 r(TP+57)=kdi*x(TP2+18); %unbinding of Rev from SNR5 to produce SNR4
231 r(TP+58)=kai*factor2*x(TP2+17)*x(TP2+7); %binding of Rev to FNR5 to
      produce FNR6
232 r(TP+59)=kdi*x(TP2+19); %unbinding of Rev from FNR6
233 r(TP+60)=(1-d)*sp*x(TP2+19); %splicing of FNR6 to produce SNR6
234 r(TP+61)=kdeg*x(TP2+19); %degradation of FNR6 to produce 6 Revs
235 r(TP+62)=(1-d)*sp*x(TP2+20); %splicing of SNR6 to produce MN and 6 Revs
236 r(TP+63)=kdeg*x(TP2+20); %degradation of SNR6 to produce 6 Revs
237 r(TP+64)=kai*factor2*x(TP2+18)*x(TP2+7); %binding of Rev to SNR5 to
      produce SNR6
238 r(TP+65)=kdi*x(TP2+20); %unbinding of Rev from SNR6 to produce SNR5
239 r(TP+66)=kai*factor2*x(TP2+19)*x(TP2+7); %binding of Rev to FNR6 to
      produce FNR7
240 r(TP+67)=kdi*x(TP2+21); %unbindnig of Rev from FNR7 to produce FNR6
241 r(TP+68)=(1-d)*sp*x(TP2+21); %splicing of FNR7 to produce SNR7
242 r(TP+69)=kdeg*x(TP2+21); %degradation of FNR7 to release 7Revs
```

```
243 r(TP+70)=kexp*x(TP2+21); %export of FNR7 to produce FC and 7 Revs in
      cytoplasm
244 r(TP+71)=(1-d)*sp*x(TP2+22); %splicing of SNR7 to produce MN and 7 Revs
245 r(TP+72)=kdeg*x(TP2+22); %degradation of SNR7 to produce 7 Revs
246 r(TP+73)=kai*factor2*x(TP2+20)*x(TP2+7); %binding of Rev to SNR6 to
      produce SNR7
247 r(TP+74)=kdi*x(TP2+22); %unbinding of Rev from SNR7 to produce SNR6
248 r(TP+75)=kexp*x(TP2+22); %export of SNR7 to produce SC and 7 Revs in
      cytoplasm
249 r(TP+76)=kai*factor2*x(TP2+21)*x(TP2+7); %binding of Rev to FNR7 to
      produce FNR8
250 r(TP+77)=kdi*x(TP2+25); %unbinding of Rev from FNR8 to produce FNR7
251 r(TP+78)=(1-d)*sp*x(TP2+25); %splicing of FNR8 to produce SNR8
252 r(TP+79)=kdeg*x(TP2+25); %degradation of FNR8 to release 8 Revs
253 r(TP+80)=kexp*x(TP2+25); %export of FNR8 to produce FC and 8 Revs in
      cytoplasm
254 r(TP+81)=(1-d)*sp*x(TP2+26); %splicing of SNR8 to produce MN and 8 Revs
255 r(TP+82)=kdeg*x(TP2+26); %degradation of SNR8 to release 8 Revs
256 r(TP+83)=kai*factor2*x(TP2+22)*x(TP2+7); %binding of Rev to SNR7 to
      produce SNR8
257 r(TP+84)=kdi*x(TP2+26); %unbinding of Rev from SNR8 to produce SNR7
258 r(TP+85)=kexp*x(TP2+26); %export of SNR8 to produce SC and 8 Revs in
      cytoplasm
259 r(TP+86)=kai*factor2*x(TP2+25)*x(TP2+7); %binding of Rev to FNR8 to
      produce FRN9
260 r(TP+87)=kdi*x(TP2+27); %unbinding of Rev from FNR9 to produce FNR8
261 r(TP+88)=(1-d)*sp*x(TP2+27); %splicing of FNR9 to produce SNR9
262 r(TP+89)=kdeg*x(TP2+27); %degradation of FNR9 to release 9 Revs
263 r(TP+90)=kexp*x(TP2+27); %export of FNR9 to produce FC and 9 Revs in
      cytoplasm
264 r(TP+91)=(1-d)*sp*x(TP2+28); %splicing of SNR9 to produce MN and 9 Revs
265 r(TP+92)=kdeg*x(TP2+28); %degradation of SNR9 to produce 9 Revs
266 r(TP+93)=kai*factor2*x(TP2+26)*x(TP2+7); %binding of Rev to SNR8 to
      produce SNR9
267 r(TP+94)=kdi*x(TP2+28); %unbinding of Rev from SNR9 to produce SNR8
268 r(TP+95)=kexp*x(TP2+28); %export of SNR9 to produce SC and 9 Revs in
```

```
    cytoplasm
269 r(TP+96)=kai*factor2*x(TP2+27)*x(TP2+7); %binding of Rev to FNR9 to
    produce FNR10
270 r(TP+97)=kdi*x(TP2+29); %unbinding of Rev from FNR10 to produce FNR9
271 r(TP+98)=(1-d)*sp*x(TP2+29); %splicing of FNR10 to produce SNR10
272 r(TP+99)=kdeg*x(TP2+29); %degradation of FNR10 to release 10 Revs
273 r(TP+100)=kexp*x(TP2+29); %export of FNR10 to produce FC and 10 Revs
    in cytoplasm
274 r(TP+101)=(1-d)*sp*x(TP2+30); %splicing of SNR10 to produce MN and 10
    Revs
275 r(TP+102)=kdeg*x(TP2+30); %degradation of SNR10 to release 10 Revs
276 r(TP+103)=kai*factor2*x(TP2+28)*x(TP2+7); %binding of Rev to SNR9 to
    produce SNR10
277 r(TP+104)=kdi*x(TP2+30); %unbinding of Rev from SNR10 to produce SNR9
278 r(TP+105)=kexp*x(TP2+30); %export of SNR10 to produce SC and 10 Revs
    in cytoplasm
279 r(TP+106)=kai*factor2*x(TP2+29)*x(TP2+7); %binding of Rev to FNR10 to
    produce FNR11
280 r(TP+107)=kdi*x(TP2+31); %unbinding of Rev from FNR11 to produce FNR10
281 r(TP+108)=(1-d)*sp*x(TP2+31); %splicing of FNR11 to produce SNR11
282 r(TP+109)=kdeg*x(TP2+31); %degradation of FNR11 to release 11 Revs
283 r(TP+110)=kexp*x(TP2+31); %export of FNR11 to produce FC and 11 Revs
    in cytoplasm
284 r(TP+111)=(1-d)*sp*x(TP2+32); %splicing of SNR11 to produce MN and 11
    Revs
285 r(TP+112)=kdeg*x(TP2+32); %degradation of SNR11 to release 11 Revs
286 r(TP+113)=kai*factor2*x(TP2+30)*x(TP2+7); %binding of Rev to SNR10 to
    produce SNR11
287 r(TP+114)=kdi*x(TP2+32); %unbinding of Rev from SNR11 to produce SNR10
288 r(TP+115)=kexp*x(TP2+32); %export of SNR11 to produce SC and 11 Revs
    in cytoplasm
289 r(TP+116)=kai*factor2*x(TP2+31)*x(TP2+7); %binding of Rev to FNR11 to
    produce FNR12
290 r(TP+117)=kdi*x(TP2+33); %unbinding of Rev from FNR12 to produce FNR11
291 r(TP+118)=(1-d)*sp*x(TP2+33); %splicing of FNR12 of produce SNR12
292 r(TP+119)=kdeg*x(TP2+33); %degradation of FNR12 to release 12 Revs
```

```
293 r(TP+120)=kexp*x(TP2+33); %export of FNR12 to produce FC and 12 Revs
        in cytoplasm
294 r(TP+121)=(1-d)*sp*x(TP2+34); %splicing of SNR12 to produce MN and 12
        Revs
295 r(TP+122)=kdeg*x(TP2+34); %degradation of SNR12 to release 12 Revs
296 r(TP+123)=kai*factor2*x(TP2+32)*x(TP2+7); %binding of Rev to SNR11 to
        produce SNR12
297 r(TP+124)=kdi*x(TP2+34); %unbinding of Rev from SNR12 to produce SNR11
298 r(TP+125)=kexp*x(TP2+34); %export of SNR12 to produce SC and 12 Revs
        in cytoplasm
299 r(TP+126)=fGag*tr*x(TP2+23); %translation of Gag protein from
        full-length mRNA in cytoplasm
300 r(TP+127)=kGag*x(TP2+35); %degradation of Gag protein
301 r(TP+128)=kbud*x(TP2+35); %export of Gag due to budding
302 r(TP+129)=2*(kbud*(x(TP2+35)))/nGagvirion; %export of 2 gmRNA due to
        budding
303 r(TP+130)=tr*0.05*x(TP2+23);%translation of GagPol from FC
304 r(TP+131)=kGag*1.1*x(TP2+36); %degradation of GagPol in cytoplasm??
305 r(TP+132)=kbud*x(TP2+36); %budding of GagPol
306 r(TP+133)=0.15*tr*x(TP2+24); %translation of gp160 from SC
307 r(TP+134)=kEnv*x(TP2+37); %degradation of gp160 in cytoplasm
308 r(TP+135)=kbud*x(TP2+37); %budding of gp160
309 CGagVif=aGagVif*x(TP2+35)*x(TP2+38);
310 r(TP+136)=0.1*tr*x(TP2+24); %translation of Vif from SC
311 r(TP+137)=kVif*x(TP2+38); %degradation of Vif
312 r(TP+138)=kbud*CGagVif;%budding of Vif
313 r(TP+139)=tr*0.27*x(TP2+24);%translation of Vpr
314 r(TP+140)=kVpr*x(TP2+39);%degradation of Vpr
315 r(TP+141)=kbud*x(TP2+39); %budding of Vpr
316 r(TP+142)=0.1*tr*x(TP2+24); % translation of Vpu
317 r(TP+143)=kVpu*x(TP2+40); % degradation of Vpu
318 r(TP+144)=kbud*x(TP2+40); %budding of Vpu
319 r(TP+145)=(0.5*tr)*x(TP2+4); %Translation of Nef
320 r(TP+146)=(kRevc*x(TP2+41)); %degradation of Nef x(TP2+41)=Nef
321 r(TP+147)=kbud*x(TP2+41); %budding of Nef
322 %Production of Rev peptides
```

```
323 r(TP+147+1:TP+147+nRev)= ps(1:nRev)*kRevc*x(TP2+5); %where
        x(TP2+5)=Revc
324 %Production of Tat Peptides
325 r(TP+147+nRev+1:TP+147+nRev+nTat)=ps(nRev+1:nRev+nTat)*kTat*x(TP2+6);
        %x(TP2+6)=Tatc
326 %Production of Nef Peptides
327 r(TP+147+nRev+nTat+1:TP+147+nRev+nTat+nNef)=ps(nRev+nTat+1:nRev+nTat+nNef)
328 *kRevc*x(TP2+41); %x(TP2+41)=Nefc
329 %Production of Gag Peptides
330 nRTN=nRev+nTat+nNef;
331 r(TP+147+nRTN+1:TP+147+nRTN+nGag)=ps(nRTN+1:nRTN+nGag)*kGag*x(TP2+35);
        % x(TP2+35)=Gag
332 %Production of GagPol Peptides
333 nRTNG=nRTN+nGag;
334 r(TP+147+nRTNG+1:TP+147+nRTNG+nGagPol)=ps(nRTNG+1:nRTNG+nGagPol)
335 *kGag*1.1*x(TP2+36); % x(TP2+36)=GagPol
336 %Production of Env Peptides
337 nRTNGP=nRTNG+nGagPol;
338 r(TP+147+nRTNGP+1:TP+147+nRTNGP+nEnv)=ps(nRTNGP+1:nRTNGP+nEnv)
339 *kEnv*x(TP2+37); % x(TP2+37)=Env
340 %Production of Vif Peptides
341 nRTNGPE=nRTNGP+nEnv;
342 r(TP+147+nRTNGPE+1:TP+147+nRTNGPE+nVif)=ps(nRTNGPE+1:nRTNGPE+nVif)
343 *kVif*x(TP2+38); %x(TP2+38)=Vif
344 %Production of Vpr Peptides
345 nRTNGPEV=nRTNGPE+nVif;
346 r(TP+147+nRTNGPEV+1:TP+147+nRTNGPEV+nVpr)=ps(nRTNGPEV+1:nRTNGPEV+nVpr)
347 *kVpr*x(TP2+39); %x(TP2+39)=Vpr
348 %Production of Vpu Peptides
349 nRTNGPEVV=nRTNGPEV+nVpr;
350 r(TP+147+nRTNGPEVV+1:TP+147+nRTNGPEVV+nVpu)=ps(nRTNGPEVV+1:nRTNGPEVV+nVpu)
351 *kVpu*x(TP2+40); % x(TP2+40)=Vpu
352 %Degradation of peptides in the cytoplasm
353 nss=TP+147+n;
354 r(nss+1:nss+n)=dPeps(:).*x(TP2+49+1:TP2+49+n); %dP = degradation of
        peptides in cytoplasm x(TP2+49+1:TP2+49+n)=CPi
```

```
NN=nss+n;
r(NN+1:NN+n,1)=g(:).*x(TP2+49+1:TP2+49+n);  %peptide supply =
    g(1:n).*(CP1:CPn)
r(NN+n+1:NN+(2*n))=dP*x(TP2+49+n+1:TP2+49+(2*n));  % peptide
    degradation where P=x(TP2+49+n+1:TP2+49+2*n);
r(NN+(2*n)+1:NN+(3*n))= bPeps(:).*x(TP2+49+n+1:TP2+49+(2*n))*x(1); %
    binding of Pi to M where x(1)=M
r(NN+(3*n)+1:NN+(4*n))= u(:).*x(TP2+49+(2*n)+1:TP2+49+(3*n)); %
    unbinding of Pi from M where x(TP2+49+2*n+1:TP2+49+3*n)=M-Pi.
r(NN+(4*n)+1:NN+(5*n))=c*x(TP2+49+n+1:TP2+49+(2*n))*x(3); % peptide
    binding to M-T where x(3)=M-T
r(NN+(5*n)+1:NN+(6*n))= q*u(:).*x(TP2+49+(3*n)+1:TP2+49+(4*n));
    %unbinding of peptide from M-T, where
    x(TP2+49+(3*n)+1:TP2+49+4*n)=M-T-Pi
r(NN+(6*n)+1:NN+(7*n))= uTv*x(TP2+49+(3*n)+1:TP2+49+(4*n)); %unbinding
    of T from M-Pi
r(NN+(7*n)+1:NN+(8*n)) = e*x(TP2+49+(2*n)+1:TP2+49+(3*n)); %egression
    of Me-Pi where x(TP2+49+2*n+1:TP2+49+3*n)=M-Pi
r(NN+(8*n)+1:NN+(9*n)) = u(:).*x(TP2+49+(4*n)+1:TP2+49+(5*n)); %
    unbinding of Pi from Me where x(TP2+49+(4*n)+1:TP2+49+5*n)= MePi
r(NN+(9*n)+1:NN+(10*n))= 0;% dMe*x(4); % where x(4) = Me
dxdt=zeros(TP2+49+(5*n),1);
dMdt1= -r(NN+(2*n)+1:NN+(3*n)) + r(NN+(3*n)+1:NN+(4*n));
dMdt2=-r(6+(4*nSelf)+1:6+(5*nSelf))+r(6+(5*nSelf)+1:6+(6*nSelf));
dxdt(1)=sum(dMdt1)+sum(dMdt2) + r(1)- r(2) -r(5) +r(6); %dM/dt
dTdt1=r(NN+(6*n)+1:NN+(7*n));
dTdt2=r(6+(8*nSelf)+1:6+(9*nSelf));
dxdt(2)=sum(dTdt1)+sum(dTdt2)+r(3)-r(4) -r(5) + r(6); %dT/dt
dMTdt1= -r(NN+(4*n)+1:NN+(5*n)) +r(NN+(5*n)+1:NN+(6*n));
dMTdt2=-r(6+(6*nSelf)+1:6+(7*nSelf))+r(6+(7*nSelf)+1:6+(8*nSelf));
dxdt(3)=sum(dMTdt1)+sum(dMTdt2) + r(5) - r(6); %dMT/dt
dMedt= r(NN+(8*n)+1:NN+(9*n))- r(NN+(9*n)+1:NN+(10*n));
dMedtSelf=r(6+(10*nSelf)+1:6+(11*nSelf))-r(6+(11*nSelf)+1:6+(12*nSelf));
dxdt(4)= sum(dMedt) +sum(dMedtSelf)-dMe*x(4);  % dMe/dt
dxdt(5:4+nSelf)=0;%r(7:6+nSelf)-r(6+nSelf+1:6+(2*nSelf))
-r(6+(2*nSelf)+1:6+(3*nSelf)); % dSelfPepCyt/dt
```

```
dxdt(4+nSelf+1:4+(2*nSelf))=r(6+(2*nSelf)+1:6+(3*nSelf))
-r(6+(3*nSelf)+1:6+(4*nSelf))-r(6+(4*nSelf)+1:6+(5*nSelf))
+r(6+(5*nSelf)+1:6+(6*nSelf))-r(6+(6*nSelf)+1:6+(7*nSelf))
+r(6+(7*nSelf)+1:6+(8*nSelf)); %dSelfER/dt
dxdt(4+(2*nSelf)+1:4+(3*nSelf))=r(6+(4*nSelf)+1:6+(5*nSelf))
-r(6+(5*nSelf)+1:6+(6*nSelf))+r(6+(8*nSelf)+1:6+(9*nSelf))
-r(6+(9*nSelf)+1:6+(10*nSelf)); %dM-SelfPep/dt
dxdt(4+(3*nSelf)+1:4+(4*nSelf))=r(6+(6*nSelf)+1:6+(7*nSelf))
-r(6+(7*nSelf)+1:6+(8*nSelf))-r(6+(8*nSelf)+1:6+(9*nSelf));
    %dM-T-SelfPep/dt
dxdt(4+(4*nSelf)+1:4+(5*nSelf))=r(6+(9*nSelf)+1:6+(10*nSelf))
-r(6+(10*nSelf)+1:6+(11*nSelf)); %dMe-SelfPep/dt
if t<8.5
    dxdt(TP2+1:TP2+4)=0;
else
dxdt(TP2+1)=r(TP+1)-r(TP+2)-r(TP+3)-r(TP+18)+r(TP+19);%FN
dxdt(TP2+2)=r(TP+2)-r(TP+4)-r(TP+5)-r(TP+24)+r(TP+25);%SN
dxdt(TP2+3)=r(TP+4)-r(TP+6)-r(TP+7)+r(TP+22)+r(TP+30)+r(TP+38)+r(TP+46)
+r(TP+54)+r(TP+62)+r(TP+71)+r(TP+81)+r(TP+91)+r(TP+101)+r(TP+111)
+r(TP+121);%MN
dxdt(TP2+4)=r(TP+6)-r(TP+8); %MC
end
if t<8.5
    dxdt(TP2+5)=-kRevc*x(TP2+5);
    dxdt(TP2+6)=-kTat*x(TP2+6);
else
dxdt(TP2+5)=r(TP+9)-r(TP+10)-r(TP+11)+r(TP+17)+(7*r(TP+70))+(7*r(TP+75))
+(8*r(TP+80))+(8*r(TP+85))+(9*r(TP+90))+(9*r(TP+95))+(10*r(TP+100))
+(10*r(TP+105))+(11*r(TP+110))+(11*r(TP+115))+(12*r(TP+120))
+(12*r(TP+125)); %Rev Cytoplasm
dxdt(TP2+6)=r(TP+12)-r(TP+13)-r(TP+14)+fRNAt*tr*fTat*x(TP2+24); %Tat
    Cytoplasm
end
if t<8.5
    dxdt(TP2+7:TP2+34)=0;
else
```

```
dxdt(TP2+7)=r(TP+11)-r(TP+15)-r(TP+17)-r(TP+18)+r(TP+19)+r(TP+21)+r(TP+22)
+r(TP+23)-r(TP+24)+r(TP+25)-r(TP+26)+r(TP+27)+(2*r(TP+29))+(2*r(TP+30))
+(2*r(TP+31))-r(TP+32)+r(TP+33)-r(TP+34)+r(TP+35)+(3*r(TP+37))+(3*r(TP+38))
+(3*r(TP+39))-r(TP+40)+r(TP+41)-r(TP+42)+r(TP+43)+(4*r(TP+45))+(4*r(TP+46))
+(4*r(TP+47))-r(TP+48)+r(TP+49)-r(TP+50)+r(TP+51)+(5*r(TP+53))+(5*r(TP+54))
+(5*r(TP+55))-r(TP+56)+r(TP+57)-r(TP+58)+r(TP+59)+(6*r(TP+61))+(6*r(TP+62))
+(6*r(TP+63))-r(TP+64)+r(TP+65)-r(TP+66)+r(TP+67)+(7*r(TP+69))+(7*r(TP+71))
+(7*r(TP+72))-r(TP+73)+r(TP+74)-r(TP+76)+r(TP+77)+(8*r(TP+79))+(8*r(TP+81))
+(8*r(TP+82))-r(TP+83)+r(TP+84)-r(TP+86)+r(TP+87)+(9*r(TP+89))+(9*r(TP+91))
+(9*r(TP+92))-r(TP+93)+r(TP+94)-r(TP+96)+r(TP+97)+(10*r(TP+99))+(10*r(TP+101))
+(10*r(TP+102))-r(TP+103)+r(TP+104)-r(TP+106)+r(TP+107)+(11*r(TP+109))
+(11*r(TP+111))+(11*r(TP+112))-r(TP+113)+r(TP+114)-r(TP+116)+r(TP+117)
+(12*r(TP+119))+(12*r(TP+121))+(12*r(TP+122))-r(TP+123)+r(TP+124);
    %Revn
dxdt(TP2+8)=r(TP+14)-r(TP+16); %Tatn
dxdt(TP2+9)=r(TP+18)-r(TP+19)-r(TP+20)-r(TP+21)-r(TP+26)+r(TP+27);
    %FNR1
dxdt(TP2+10)=r(TP+20)-r(TP+22)-r(TP+23)+r(TP+24)-r(TP+25)-r(TP+32)+r(TP+33);
    %SNR1
dxdt(TP2+11)=r(TP+26)-r(TP+27)-r(TP+28)-r(TP+29)-r(TP+34)+r(TP+35);
    %FNR2
dxdt(TP2+12)=r(TP+28)-r(TP+30)-r(TP+31)+r(TP+32)-r(TP+33)-r(TP+40)+r(TP+41);
    %SNR2
dxdt(TP2+13)=r(TP+34)-r(TP+35)-r(TP+36)-r(TP+37)-r(TP+42)+r(TP+43);
    %FNR3
dxdt(TP2+14)=r(TP+36)-r(TP+38)-r(TP+39)+r(TP+40)-r(TP+41)-r(TP+48)+r(TP+49);
    %SNR3
dxdt(TP2+15)=r(TP+42)-r(TP+43)-r(TP+44)-r(TP+45)-r(TP+50)+r(TP+51);
    %FNR4
dxdt(TP2+16)=r(TP+44)-r(TP+46)-r(TP+47)+r(TP+48)-r(TP+49)-r(TP+56)+r(TP+57);
    %SNR4
dxdt(TP2+17)=r(TP+50)-r(TP+51)-r(TP+52)-r(TP+53)-r(TP+58)+r(TP+59);
    %FNR5
dxdt(TP2+18)=r(TP+52)-r(TP+54)-r(TP+55)+r(TP+56)-r(TP+57)-r(TP+64)+r(TP+65);
    %SNR5
dxdt(TP2+19)=r(TP+58)-r(TP+59)-r(TP+60)-r(TP+61)-r(TP+66)+r(TP+67);
```

```
    %FNR6
dxdt(TP2+20)=r(TP+60)-r(TP+62)-r(TP+63)+r(TP+64)-r(TP+65)-r(TP+73)+r(TP+74);
    %SNR6
dxdt(TP2+21)=r(TP+66)-r(TP+67)-r(TP+68)-r(TP+69)-r(TP+70)-r(TP+76)+r(TP+77);
    %FNR7
dxdt(TP2+22)=r(TP+68)-r(TP+71)-r(TP+72)+r(TP+73)-r(TP+74)-r(TP+75)-r(TP+83)
+r(TP+84); %SNR7
dxdt(TP2+23)=r(TP+70)+r(TP+80)+r(TP+90)+r(TP+100)+r(TP+110)+r(TP+120)
-kdeg*x(TP2+23)-r(TP+129); %FC
dxdt(TP2+24)=r(TP+75)+r(TP+85)+r(TP+95)+r(TP+105)+r(TP+115)+r(TP+125)
-kdeg*x(TP2+24); %SC
dxdt(TP2+25)=r(TP+76)-r(TP+77)-r(TP+78)-r(TP+79)-r(TP+80)-r(TP+86)+r(TP+87);
    %FNR8
dxdt(TP2+26)=r(TP+78)-r(TP+81)-r(TP+82)+r(TP+83)-r(TP+84)-r(TP+85)-r(TP+93)
+r(TP+94); %SNR8
dxdt(TP2+27)=r(TP+86)-r(TP+87)-r(TP+88)-r(TP+89)-r(TP+90)-r(TP+96)+r(TP+97);
    %FNR9
dxdt(TP2+28)=r(TP+88)-r(TP+91)-r(TP+92)+r(TP+93)-r(TP+94)-r(TP+95)-r(TP+103)
+r(TP+104); %SNR9
dxdt(TP2+29)=r(TP+96)-r(TP+97)-r(TP+98)-r(TP+99)-r(TP+100)-r(TP+106)+r(TP+107);
    %FNR10
dxdt(TP2+30)=r(TP+98)-r(TP+101)-r(TP+102)+r(TP+103)-r(TP+104)-r(TP+105)
-r(TP+113)+r(TP+114); %SNR10
dxdt(TP2+31)=r(TP+106)-r(TP+107)-r(TP+108)-r(TP+109)-r(TP+110)-r(TP+116)
+r(TP+117); %FNR11
dxdt(TP2+32)=r(TP+108)-r(TP+111)-r(TP+112)+r(TP+113)-r(TP+114)-r(TP+115)
-r(TP+123)+r(TP+124); %SNR11
dxdt(TP2+33)=r(TP+116)-r(TP+117)-r(TP+118)-r(TP+119)-r(TP+120); %FNR12
dxdt(TP2+34)=r(TP+118)-r(TP+121)-r(TP+122)+r(TP+123)-r(TP+124)-r(TP+125);
    %SNR12
end
if t<8.5
    dxdt(TP2+35)=-kGag*x(TP2+35);
    dxdt(TP2+36)=-kGag*1.1*x(TP2+36);
    dxdt(TP2+37)=-kEnv*x(TP2+37);
    dxdt(TP2+38)=-kVif*x(TP2+38);
```

```
469         dxdt(TP2+39)=-kVpr*x(TP2+39);
470         dxdt(TP2+40)=-kVpu*x(TP2+40);
471         dxdt(TP2+41)=-kRevc*x(TP2+41);
472         dxdt(TP2+42:49)=0;
473 else
474 dxdt(TP2+35)=r(TP+126)-r(TP+127)-r(TP+128); %Gag
475 dxdt(TP2+36)=r(TP+130)-r(TP+131)-r(TP+132); %GagPol
476 dxdt(TP2+37)=r(TP+133)-r(TP+134)-r(TP+135); %Env
477 dxdt(TP2+38)=r(TP+136)-r(TP+137)-r(TP+138); %Vif
478 dxdt(TP2+39)=r(TP+139)-r(TP+140)-r(TP+141); %Vpr
479 dxdt(TP2+40)=r(TP+142)-r(TP+143)-r(TP+144); %Vpu
480 dxdt(TP2+41)=r(TP+145)-r(TP+146)-r(TP+147); %Nef
481 dxdt(TP2+42)=(kbud*(x(TP2+35)))/nGagvirion; %Virion'
482 dxdt(TP2+43)=r(TP+128);%/((kbud*(x(35)))/nGagvirion); %GagVirion
483 dxdt(TP2+44)=r(TP+132);%/((kbud*(x(35)))/nGagvirion); %GagPolVirion
484 dxdt(TP2+45)=r(TP+135);%/((kbud*(x(35)))/nGagvirion); %EnvVirion
485 dxdt(TP2+46)=r(TP+138);%/((kbud*(x(35)))/nGagvirion); %VifVirion
486 dxdt(TP2+47)=r(TP+141);%/((kbud*(x(35)))/nGagvirion); %VprVirion
487 dxdt(TP2+48)=r(TP+144);%/((kbud*(x(35)))/nGagvirion); %VpuVirion
488 dxdt(TP2+49)=r(TP+147);%/((kbud*(x(35)))/nGagvirion); %NefVirion
489 end
490 dxdt(TP2+49+1:TP2+49+n)=r(TP+147+1:TP+147+n)-r(NN+1:NN+n)-r(nss+1:nss+n);
      % dCPi/dt = production-suppply to ER - degradation in cytoplasm
491 dxdt(TP2+49+n+1:TP2+49+(2*n))=r(NN+1:NN+n)-r(NN+n+1:NN+(2*n))
492 -r(NN+(2*n)+1:NN+(3*n)) + r(NN+(3*n)+1:NN+(4*n))
493 - r(NN+(4*n)+1:NN+(5*n)) +r(NN+(5*n)+1:NN+(6*n)); %dPi/dt
494 dxdt(TP2+49+(2*n)+1:TP2+49+(3*n))= r(NN+(2*n)+1:NN+(3*n))
495 - r(NN+(3*n)+1:NN+(4*n)) + r(NN+(6*n)+1:NN+(7*n))
496  - r(NN+(7*n)+1:NN+(8*n)); %dMPi/dt
497 dxdt(TP2+49+(3*n)+1:TP2+49+(4*n))= r(NN+(4*n)+1:NN+(5*n))
498  - r(NN+(5*n)+1:NN+(6*n)) - r(NN+(6*n)+1:NN+(7*n)); % dMTPi/dt
499 dxdt(TP2+49+(4*n)+1:TP2+49+(5*n))= r(NN+(7*n)+1:NN+(8*n)) -
      r(NN+(8*n)+1:NN+(9*n)); % dMePi/dt
500 end
```

Bibliography

[1] Iwasaki, A. & Medzhitov, R. Control of adaptive immunity by the innate immune system. *Nat Immunol* **16**, 343–353 (2015). URL http://dx.doi.org/10.1038/ni.3123http://10.0.4.14/ni.3123.

[2] Medzhitov, R. & Janeway, C. A. Innate immunity: impact on the adaptive immune response. *Current Opinion in Immunology* **9**, 4–9 (1997). URL http://www.sciencedirect.com/science/article/pii/S0952791597801525.

[3] Hoebe, K., Janssen, E. & Beutler, B. The interface between innate and adaptive immunity. *Nat Immunol* **5**, 971–974 (2004). URL http://dx.doi.org/10.1038/ni1004-971.

[4] et al. Alberts B, Johnson A, Lewis J. Molecular Biology of the Cell. chap. Innate Imm (Garland Science, New York, 2002), 4 edn. URL https://www.ncbi.nlm.nih.gov/books/NBK26846/.

[5] SA, F. Immunology and Evolution of Infectious Disease. In *Immunology and Evolution of Infectious Disease.*, chap. 6 (Princeton University Press, Princeton (NJ), 2002). URL https://www.ncbi.nlm.nih.gov/books/NBK2386/.

[6] Demas, G. & Nelson, R. *Ecoimmunology* (Oxford University Press, New York, 2012).

[7] Fortier, M.-H. *et al.* The MHC class I peptide repertoire is molded by the transcriptome. *The Journal of Experimental Medicine* **205**, 595–610 (2008). URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2275383/.

[8] Kotturi, M. F. *et al.* Naive Precursor Frequencies and MHC Binding Rather Than the Degree of Epitope Diversity Shape CD8(+) T Cell Immunodominance(). *Journal of Immunology (Baltimore, Md. : 1950)* **181**, 2124–2133 (2008). URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3319690/.

[9] Wu, L. C., Tuot, D. S., Lyons, D. S., Garcia, K. C. & Davis, M. M. Two-step binding mechanism for T-cell receptor recognition of peptide-MHC. *Nature* **418**, 552–556 (2002). URL http://dx.doi.org/10.1038/nature00920http://www.nature.com/nature/journal/v418/n6897/suppinfo/nature00920{_}S1.html.

[10] Burrows, S. R. *et al.* Hard wiring of T cell receptor specificity for the major histocompatibility complex is underpinned by TCR adaptability. *Proceedings of the National Academy of Sciences* **107**, 10608–10613 (2010). URL http://www.pnas.org/content/107/23/10608.abstract.

[11] Monks, C. R. F., Freiberg, B. A., Kupfer, H., Sciaky, N. & Kupfer, A. Three-dimensional segregation of supramolecular activation clusters in T cells. *Nature* **395**, 82–86 (1998). URL http://dx.doi.org/10.1038/25764.

[12] Irvine, D. J., Purbhoo, M. A., Krogsgaard, M. & Davis, M. M. Direct observation of ligand recognition by T cells. *Nature* **419**, 845–849 (2002). URL http://dx.doi.org/10.1038/nature01076http://www.nature.com/nature/journal/v419/n6909/suppinfo/nature01076{_}S1.html.

[13] Lavoie, P. M., Dumont, A. R., McGrath, H., Kernaleguen, A.-E. & Sékaly, R.-P. Delayed expansion of a restricted T cell repertoire by low-density TCR ligands. *International Immunology* **17**, 931–941 (2005). URL http://dx.doi.org/10.1093/intimm/dxh273.

[14] Morris, G. P. & Allen, P. M. How the TCR balances sensitivity and specificity for the recognition of self and pathogens. *Nat Immunol* **13**, 121–128 (2012). URL http://dx.doi.org/10.1038/ni.2190.

[15] Yewdell, J. W., Antón, L. C. & Bennink, J. R. Defective ribosomal products (DRiPs): a major source of antigenic peptides for MHC class I molecules? *The Journal of Immunology* **157**, 1823 LP – 1826 (1996). URL http://www.jimmunol.org/content/157/5/1823.abstract.

[16] Bourdetsky, D., Schmelzer, C. E. H. & Admon, A. The nature and extent of contributions by defective ribosome products to the HLA peptidome. *Proceedings of the National Academy of Sciences* **111**, E1591–E1599 (2014). URL http://www.pnas.org/content/111/16/E1591.abstract.

[17] Van Hateren, A. *et al.* The cell biology of major histocompatibility complex class I assembly: towards a molecular understanding. *Tissue Antigens* **76**, 259–275 (2010). URL http://dx.doi.org/10.1111/j.1399-0039.2010.01550.x.

[18] Boulanger, D. S. M. *et al.* Absence of Tapasin Alters Immunodominance against a Lymphocytic Choriomeningitis Virus Polytope. *The Journal of Immunology* **184**, 73 LP – 83 (2009). URL http://www.jimmunol.org/content/184/1/73.abstract.

[19] Morozov, G. I. *et al.* Interaction of TAPBPR, a tapasin homolog, with MHC-I molecules promotes peptide editing. *Proceedings of the National Academy of Sciences* **113**, E1006–E1015 (2016). URL http://www.pnas.org/content/113/8/E1006.abstract.

[20] Hermann, C. *et al.* TAPBPR alters MHC class I peptide presentation by functioning as a peptide exchange catalyst. *eLife* **4**, e09617 (2015). URL https://doi.org/10.7554/eLife.09617.

[21] Croft, N. P. *et al.* Kinetics of Antigen Expression and Epitope Presentation during Virus Infection. *PLoS Pathogens* **9** (2013).

[22] Dalchau, N. *et al.* A peptide filtering relation quantifies MHC class I peptide optimization. *PLoS Computational Biology* **7** (2011).

[23] Livingston, B. *et al.* A Rational Strategy to Design Multiepitope Immunogens Based on Multiple Th Lymphocyte Epitopes. *The Journal of Immunology* **168**, 5499 LP – 5506 (2002). URL http://www.jimmunol.org/content/168/11/5499.abstract.

[24] Jost, S. *et al.* A Patient with HIV-1 Superinfection. *New England Journal of Medicine* **347**, 731–736 (2002). URL http://dx.doi.org/10.1056/NEJMoa020263.

[25] Toes, R. E. M. *et al.* Protective anti-tumor immunity induced by vaccination with recombinant adenoviruses encoding multiple tumor-associated cytotoxic T lymphocyte epitopes in a string-of-beadsfashion. *Proceedings of the National Academy of Sciences* **94**, 14660–14665 (1997). URL http://www.pnas.org/content/94/26/14660.abstract.

[26] Cantor, J. R. *et al.* Therapeutic enzyme deimmunization by combinatorial T-cell epitope removal using neutral drift. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 1272–1277 (2011). URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3029727/.

[27] Fischer, H. P. Mathematical Modeling of Complex Biological Systems: From Parts Lists to Understanding Systems Behavior. *Alcohol Research & Health* **31**, 49–59 (2008). URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3860444/.

[28] Chance, M. R. *et al.* High-Throughput Computational and Experimental Techniques in Structural Genomics. *Genome Research* **14**, 2145–2154 (2004). URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC528931/.

[29] Duggan, D. J., Bittner, M., Chen, Y., Meltzer, P. & Trent, J. M. Expression profiling using cDNA microarrays. *Nat Genet* (1999).

[30] Hondowicz, B. D. *et al.* Discovery of T Cell Antigens by High-Throughput Screening of Synthetic Minigene Libraries. *PLOS ONE* **7**, e29949 (2012). URL https://doi.org/10.1371/journal.pone.0029949.

[31] Harndahl, M., Rasmussen, M., Roder, G. & Buus, S. Real-time, High-Throughput Measurements of Peptide-MHC-I Dissociation Using a Scintillation Proximity Assay. *Journal of immunological methods* **374**, 5–12 (2011). URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4341823/.

[32] Clayton, R. H. & Panfilov, A. V. A guide to modelling cardiac electrical activity in anatomically detailed ventricles. *Progress in Biophysics and Molecular Biology* **96**, 19–43 (2008). URL http://www.sciencedirect.com/science/article/pii/S0079610707000454.

[33] Mayer, H., Zaenker, K. S. & an der Heiden, U. A basic mathematical model of the immune response. *Chaos: An Interdisciplinary Journal of Nonlinear Science* **5**, 155–161 (1995). URL http://dx.doi.org/10.1063/1.166098.

[34] Wodarz, D. & Nowak, M. A. Mathematical models of HIV pathogenesis and treatment. *BioEssays* **24**, 1178–1187 (2002). URL http://dx.doi.org/10.1002/bies.10196.

[35] McKeithan, T. W. Kinetic proofreading in T-cell receptor signal transduction. *Proceedings of the National Academy of Sciences* **92**, 5042–5046 (1995). URL http://www.pnas.org/content/92/11/5042.abstract.

[36] Hopfield, J. J. Kinetic Proofreading: A New Mechanism for Reducing Errors in Biosynthetic Processes Requiring High Specificity. *Proceedings of the National Academy of Sciences* **71**, 4135–4139 (1974). URL http://www.pnas.org/content/71/10/4135.abstract.

[37] Ninio, J. Kinetic amplification of enzyme discrimination. *Biochimie* **57**, 587–595 (1975). URL http://www.sciencedirect.com/science/article/pii/S0300908475801398.

[38] Dushek, O., Das, R. & Coombs, D. A Role for Rebinding in Rapid and Reliable T Cell Responses to Antigen. *PLOS Computational Biology* **5**, e1000578 (2009). URL https://doi.org/10.1371/journal.pcbi.1000578.

[39] van den Berg, H. A., Burroughs, N. J. & Rand, D. A. Quantifying the strength of ligand antagonism in TCR triggering. *Bulletin of Mathematical Biology* **64**, 781–808 (2002). URL https://doi.org/10.1006/bulm.2002.0302.

[40] Coombs, D., Kalergis, A. M., Nathenson, S. G., Wofsy, C. & Goldstein, B. Activated TCRs remain marked for internalization after dissociation from pMHC. *Nat Immunol* **3**, 926–931 (2002). URL http://dx.doi.org/10.1038/ni838http://www.nature.com/ni/journal/v3/n10/suppinfo/ni838{_}S1.html.

[41] Feinerman, O., Germain, R. N. & Altan-Bonnet, G. Quantitative challenges in understanding ligand discrimination by $\alpha\beta$ T cells. *Molecular immunology* **45**, 619–631 (2008). URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2131735/.

[42] Stefanova, I. *et al.* TCR ligand discrimination is enforced by competing ERK positive and SHP-1 negative feedback pathways. *Nat Immunol* **4**, 248–254 (2003). URL http://dx.doi.org/10.1038/ni895http://www.nature.com/ni/journal/v4/n3/suppinfo/ni895{_}S1.html.

[43] Altan-Bonnet, G. & Germain, R. N. Modeling T Cell Antigen Discrimination Based on Feedback Control of Digital ERK Responses. *PLOS Biology* **3**, e356 (2005). URL https://doi.org/10.1371/journal.pbio.0030356.

[44] Wylie, D. C., Das, J. & Chakraborty, A. K. Sensitivity of T cells to antigen and antagonism emerges from differential regulation of the same molecular signaling module. *Proceedings of the National Academy of Sciences* **104**, 5533–5538 (2007). URL http://www.pnas.org/content/104/13/5533.abstract.

[45] Feinerman, O., Veiga, J., Dorfman, J. R., Germain, R. N. & Altan-Bonnet, G. Variability and Robustness in T Cell Activation from Regulated Heterogeneity in Protein Levels. *Science (New York, N.Y.)* **321**, 1081–1084 (2008). URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2673522/.

[46] Casal, A., Sumen, C., Reddy, T. E., Alber, M. S. & Lee, P. P. Agent-based modeling of the context dependency in T cell recognition. *Journal of Theoretical Biology* **236**, 376–391 (2005). URL http://www.sciencedirect.com/science/article/pii/S0022519305001281.

[47] Choudhuri, K. & van der Merwe, P. A. Molecular mechanisms involved in T cell receptor triggering. *Seminars in Immunology* **19**, 255–261 (2007). URL http://www.sciencedirect.com/science/article/pii/S1044532307000590.

[48] Choudhuri, K., Kearney, A., Bakker, T. R. & van der Merwe, P. A. Immunology: How Do T Cells Recognize Antigen? *Current Biology* **15**, R382–R385 (2005). URL http://www.sciencedirect.com/science/article/pii/S0960982205004884.

[49] Ma, Z., Janmey, P. A. & Finkel, T. H. The receptor deformation model of TCR triggering. *The FASEB Journal* **22**, 1002–1008 (2008). URL http://www.fasebj.org/content/22/4/1002.abstract.

[50] Xu, C. *et al.* Regulation of T cell Receptor Activation by Dynamic Membrane Binding of the CD3ε Cytoplasmic Tyrosine-Based Motif. *Cell* **135**, 702–713 (2008). URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2597348/.

[51] Gil, D., Schamel, W. W. A., Montoya, M., Sánchez-Madrid, F. & Alarcón, B. Recruitment of Nck by CD3 Reveals a Ligand-Induced Conformational Change Essential for T Cell Receptor Signaling and Synapse Formation. *Cell* **109**, 901–912 (2002). URL http://www.sciencedirect.com/science/article/pii/S0092867402007997.

[52] Aivazian, D. & Stern, L. J. Phosphorylation of T cell receptor [zeta] is regulated by a lipid dependent folding transition. *Nat Struct Mol Biol* **7**, 1023–1026 (2000). URL http://dx.doi.org/10.1038/80930.

[53] Cochran, J. R., Cameron, T. O. & Stern, L. J. The Relationship of MHC-Peptide Binding and T Cell Activation Probed Using Chemically Defined MHC Class II Oligomers. *Immunity* **12**, 241–250 (2000). URL http://www.sciencedirect.com/science/article/pii/S1074761300801776.

[54] Krogsgaard, M. *et al.* Agonist/endogenous peptide-MHC heterodimers drive T cell activation and sensitivity. *Nature* **434**, 238–243 (2005). URL http://dx.doi.org/10.1038/nature03391http://www.nature.com/nature/journal/v434/n7030/suppinfo/nature03391{_}S1.html.

[55] Davis, S. J. & van der Merwe, P. The structure and ligand interactions of CD2: implications for T-cell function. *Immunology Today* **17**, 177–187 (1996). URL http://www.sciencedirect.com/science/article/pii/0167569996806177.

[56] Burroughs, N. J., Lazic, Z. & van der Merwe, P. A. Ligand Detection and Discrimination by Spatial Relocalization: A Kinase-Phosphatase Segregation Model of TCR Activation. *Biophysical Journal* **91**, 1619–1629 (2006). URL http://www.sciencedirect.com/science/article/pii/S0006349506718768.

[57] Coombs, D., Dushek, O. & van der Merwe, P. A. A Review of Mathematical Models for T Cell Receptor Triggering and Antigen Discrimination BT - Mathematical Models and Immune Cell Biology. 25–45 (Springer New York, New York, NY, 2011). URL https://doi.org/10.1007/978-1-4419-7725-0{_}2.

[58] Lundegaard, C., Lund, O. & Nielsen, M. Predictions versus high-throughput experiments in T-cell epitope discovery: competition or synergy? *Expert review of vaccines* **11**, 43–54 (2012). NIHMS150003.

[59] Yewdell, J. W. Confronting Complexity: Real-World Immunodominance in Antiviral CD8+ T Cell Responses. *Immunity* **25**, 533–543

(2006). URL http://www.sciencedirect.com/science/article/pii/S1074761306004390.

[60] Kim, Y. *et al.* Immune epitope database analysis resource. *Nucl. Acids Res.* 1–6 (2012).

[61] Parker, K. C., Bednarek, M. A. & Coligan, J. E. Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains. *Journal of immunology (Baltimore, Md. : 1950)* **152**, 163–75 (1994).

[62] Andreatta, M. & Nielsen, M. Gapped sequence alignment using artificial neural networks: application to the MHC class I system. *Bioinformatics* **32**, 511–517 (2016). URL http://dx.doi.org/10.1093/bioinformatics/btv639.

[63] Farrell, D. *et al.* Integrated computational prediction and experimental validation identifies promiscuous T cell epitopes in the proteome of Mycobacterium bovis. *Microbial Genomics* **2**, e000071 (2016). URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5320590/.

[64] Zhang, L. *et al.* TEPITOPEpan: Extending TEPITOPE for Peptide Binding Prediction Covering over 700 HLA-DR Molecules. *PLOS ONE* **7**, e30483 (2012). URL https://doi.org/10.1371/journal.pone.0030483.

[65] Nielsen, M. *et al.* Quantitative Predictions of Peptide Binding to Any HLA-DR Molecule of Known Sequence: NetMHCIIpan. *PLOS Computational Biology* **4**, e1000107 (2008). URL https://doi.org/10.1371/journal.pcbi.1000107.

[66] Chang, S. T., Linderman, J. J. & Kirschner, D. E. Multiple mechanisms allow Mycobacterium tuberculosis to continuously inhibit MHC class II-mediated antigen presentation by macrophages. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 4530–4535 (2005). URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC555518/.

[67] Moreno, C., Mehlert, A. & Lamb, J. The inhibitory effects of mycobacterial lipoarabinomannan and polysaccharides upon polyclonal and monoclonal human T cell proliferation. *Clinical and Experimental Immunology* **74**, 206–210 (1988). URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1541790/.

[68] Hmama, Z., Gabathuler, R., Jefferies, W. A., de Jong, G. & Reiner, N. E. Attenuation of HLA-DR Expression by Mononuclear Phagocytes Infected with Mycobacterium tuberculosis Is Related to Intracellular Sequestration of Immature Class II Heterodimers. *The Journal of Immunology* **161**, 4882 LP – 4893 (1998). URL http://www.jimmunol.org/content/161/9/4882.abstract.

[69] Noss, E. H., Harding, C. V. & Boom, W. Mycobacterium tuberculosis Inhibits MHC Class II Antigen Processing in Murine Bone Marrow Macrophages. *Cellular Immunology* **201**, 63–74 (2000). URL http://www.sciencedirect.com/science/article/pii/S0008874900916332.

[70] Howarth, M., Williams, A., Tolstrup, A. B. & Elliott, T. Tapasin enhances MHC class I peptide presentation according to peptide half-life. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 11737–11742 (2004). URL http://www.pnas.org/content/101/32/11737.abstract.

[71] Williams, A. P., Peh, C. A., Purcell, A. W., McCluskey, J. & Elliott, T. Optimization of the MHC Class I Peptide Cargo Is Dependent on Tapasin. *Immunity* **16**, 509–520 (2002). URL http://www.sciencedirect.com/science/article/pii/S1074761302003047.

[72] Feinberg, M. & Horn, F. J. M. Chemical mechanism structure and the coincidence of the stoichiometric and kinetic subspaces. *Archive for Rational Mechanics and Analysis* **66**, 83–97 (1977). URL https://doi.org/10.1007/BF00250853.

[73] Horn, F. & Jackson, R. General mass action kinetics. *Archive for Rational Mechanics and Analysis* **47**, 81–116 (1972). URL https://doi.org/10.1007/BF00251225.

[74] Higham, D. J. Modeling and Simulating Chemical Reactions. *Society for Industrial and Applied MathematicsVol* **50**, 347–368 (2008).

[75] Fleri, W. *et al.* The Immune Epitope Database and Analysis Resource in Epitope Discovery and Synthetic Vaccine Design. *Frontiers in Immunology* **8**, 278 (2017).

[76] Moutaftsi, M. *et al.* A consensus epitope prediction approach identifies the breadth of murine T(CD8+)-cell responses to vaccinia virus. *Nature Biotechnology* **24**, 817–9 (2006).

[77] Nielsen, M. *et al.* Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. *Protein Sci* **12** (2003). URL http://dx.doi.org/10.1110/ps.0239403.

[78] Lundegaard, C. *et al.* NetMHC-3.0: Accurate web accessible predictions of Human, Mouse, and Monkey MHC class I affinities for peptides of length 8-11. *NAR* **36**, W509–512 (2008).

[79] Peters, B., Tong, W., Sidney, J., Sette, A. & Weng, Z. Examining the independent binding assumption for binding of peptide epitopes to MHC-I molecules. *Bioinformatics* **19** (2003). URL http://dx.doi.org/10.1093/bioinformatics/btg247.

[80] Sidney, J. *et al.* Quantitative peptide binding motifs for 19 human and mouse MHC class i molecules derived using positional scanning combinatorial peptide libraries. *Immunome Research* **4** (2008).

[81] Lam, L. & Suen, S. Y. Application of majority voting to pattern recognition: an analysis of its behavior and performance. *IEEE Trans. Syst. Man Cybern.* **27**, 553–568 (1997).

[82] Dayhoff, J. E. & DeLeo, J. M. Artificial Neural Networks Opening the Black Box. *American Cancer Society* 1615–1635 (2001).

[83] Gulukota, K., Sidney, J., Sette, A. & C.DeLisi. Two complementary methods for predicting peptides binding major histocompatibility complex molecules. *Journal of Molecular Biology* **267**, 1258–1267 (1997).

[84] M.R. Segal, M. R., Cummings, M. P. & Hubbard, A. E. Relating amino acid sequence to phenotype: analysis of peptide-binding data. *Biometrics* **57**, 632–642 (2001).

[85] Doytchinova, I. A., Blythe, M. J. & Flower, D. R. Additive method for the prediction of proteinpeptide binding afnity. Application to the MHC class I molecule HLA-A*0201. *J. Proteome Res.* **1**, 263–272 (2002).

[86] Rosendahl Huber, S., van Beek, J., de Jonge, J., Luytjes, W. & van Baarle, D. T Cell Responses to Viral Infections Opportunities for Peptide Vaccination. *Frontiers in Immunology* **5**, 171 (2014). URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3997009/`.

[87] Rosario, M. *et al.* Long peptides induce polyfunctional T cells against conserved regions of HIV-1 with superior breadth to single-gene vaccines in macaques. *European Journal of Immunology* **40**, 1973–1984 (2010). URL `http://dx.doi.org/10.1002/eji.201040344`.

[88] Feltkamp, M. C. W., Vierboom, M. P. M., Kast, W. & Melief, C. J. M. Efficient MHC class I-peptide binding is required but does not ensure MHC class I-restricted immunogenicity. *Molecular Immunology* **31**, 1391–1401 (1994). URL `http://www.sciencedirect.com/science/article/pii/0161589094901554`.

[89] Ochoa-Garay, J., McKinney, D. M., Kochounian, H. H. & McMillan, M. The ability of peptides to induce cytotoxic T cells in vitro does not strongly correlate with their affinity for the H-2Ld molecule: Implications

for vaccine design and immunotherapy. *Molecular Immunology* **34**, 273–281 (1997). URL http://www.sciencedirect.com/science/article/pii/S0161589097000199.

[90] Schellens, I. M. M. *et al.* Comprehensive Analysis of the Naturally Processed Peptide Repertoire: Differences between HLA-A and B in the Immunopeptidome. *PLOS ONE* **10**, e0136417 (2015). URL https://doi.org/10.1371/journal.pone.0136417.

[91] Bassani-Sternberg, M., Pletscher-Frankild, S., Jensen, L. J. & Mann, M. Mass Spectrometry of Human Leukocyte Antigen Class I Peptidomes Reveals Strong Effects of Protein Abundance and Turnover on Antigen Presentation. *Molecular & Cellular Proteomics : MCP* **14**, 658–673 (2015). URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4349985/.

[92] Milner, E., Barnea, E., Beer, I. & Admon, A. The Turnover Kinetics of Major Histocompatibility Complex Peptides of Human Cancer Cells*. *Molecular & Cellular Proteomics* **5**, 357–365 (2006). URL http://www.mcponline.org/content/5/2/357.abstract.

[93] Pollard, K. M., Cauvi, D. M., Toomey, C. B., Morris, K. V. & Kono, D. H. Interferon-γ and Systemic Autoimmunity. *Discovery medicine* **16**, 123–131 (2013). URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3934799/.

[94] Lev, A. *et al.* Compartmentalized MHC class I antigen processing enhances immunosurveillance by circumventing the law of mass action. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 6964–6969 (2010). URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2872426/.

[95] Neijssen, J. *et al.* Cross-presentation by intercellular peptide transfer through gap junctions. *Nature* **434**, 83–88 (2005). URL http://dx.doi.org/10.

1038/nature03290http://www.nature.com/nature/journal/v434/n7029/suppinfo/nature03290{_}S1.html.

[96] Roberts, C. & Casella, G. *Monte Carlo Statistical Methods* (Springer Verlag, 1999).

[97] Dudek, N. L. *et al.* Constitutive and Inflammatory Immunopeptidome of Pancreatic β-Cells. *Diabetes* **61**, 3018–3025 (2012). URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3478525/.

[98] Kaplan, D. H. *et al.* Demonstration of an interferon γ-dependent tumor surveillance system in immunocompetent mice. *Proceedings of the National Academy of Sciences* **95**, 7556–7561 (1998). URL http://www.pnas.org/content/95/13/7556.abstract.

[99] Henrickson, S. E. *et al.* T cell sensing of antigen dose governs interactive behavior with dendritic cells and sets a threshold for T cell activation. *Nat Immunol* **9**, 282–291 (2008). URL http://dx.doi.org/10.1038/ni1559http://www.nature.com/ni/journal/v9/n3/suppinfo/ni1559{_}S1.html.

[100] Meissner, J. Nucleotide sequences and further characterization of human papillomavirus DNA present in the CaSki, SiHa and HeLa cervical carcinoma cell lines. *J. Gen. Virol.* **80**, 1725–1733 (1999).

[101] Bruni, L. *et al.* Global estimates of human papillomavirus vaccination coverage by region and income level: a pooled analysis. *The Lancet Global Health* **4**, e453–e463 (2016). URL http://www.sciencedirect.com/science/article/pii/S2214109X16300997.

[102] Nanni, P. *et al.* Combined Allogeneic Tumor Cell Vaccination and Systemic Interleukin 12 Prevents Mammary Carcinogenesis in HER-2/neu Transgenic Mice. *The Journal of Experimental Medicine* **194**, 1195 LP – 1206 (2001). URL http://jem.rupress.org/content/194/9/1195.abstract.

[103] Srivatsan, S. *et al.* Allogeneic tumor cell vaccines: The promise and limitations in clinical trials. *Human Vaccines & Immunotherapeutics* **10**, 52–63 (2014). URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4181031/.

[104] Guo, C. *et al.* Therapeutic Cancer Vaccines: Past, Present and Future. *Advances in cancer research* **119**, 421–475 (2013). URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3721379/.

[105] Slingluff, C. L. *et al.* Helper T-Cell Responses and Clinical Activity of a Melanoma Vaccine With Multiple Peptides From MAGE and Melanocytic Differentiation Antigens. *Journal of Clinical Oncology* **26**, 4973–4980 (2008). URL https://doi.org/10.1200/JCO.2008.17.3161.

[106] Rosenberg, S. A., Yang, J. C. & Restifo, N. P. Cancer immunotherapy: moving beyond current vaccines. *Nature medicine* **10**, 909–915 (2004). URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1435696/.

[107] Boisvert, F.-M. *et al.* A Quantitative Spatial Proteomics Analysis of Proteome Turnover in Human Cells. *Molecular & Cellular Proteomics : MCP* **11**, M111.011429 (2012). URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3316722/.

[108] Nagaraj, N. *et al.* Deep proteome and transcriptome mapping of a human cancer cell line. *Molecular Systems Biology* **7**, 548 (2011). URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3261714/.

[109] Tenzer, S. *et al.* Modeling the MHC class I pathway by combining predictions of proteasomal cleavage, TAP transport and MHC class I binding. *Cellular and Molecular Life Sciences* **62**, 1025–1037 (2005).

[110] Cascio, P., Hilton, C., Kisselev, A. F., Rock, K. L. & Goldberg, A. L. 26S proteasomes and immunoproteasomes produce mainly N-extended versions of an antigenic peptide. *EMBO Journal* **20**, 2357–2366 (2001).

[111] Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modeling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012). URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3320027/.

[112] Boegel, S., Löwer, M., Bukur, T., Sahin, U. & Castle, J. C. A catalog of HLA type, HLA expression, and neo-epitope candidates in human cancer cell lines. *OncoImmunology* **3**, e954893 (2014). URL http://dx.doi.org/10.4161/21624011.2014.954893.

[113] Forbes, S. A. *et al.* COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Research* **39**, D945–D950 (2011). URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3013785/.

[114] Hoof, I., van Baarle, D., Hildebrand, W. H. & Kemir, C. Proteome Sampling by the HLA Class I Antigen Processing Pathway. *PLOS Computational Biology* **8**, e1002517 (2012). URL https://doi.org/10.1371/journal.pcbi.1002517.

[115] Kloetzel, P.-M. Antigen processing by the proteasome . *Nat Rev Mol Cell Biol* **2**, 179–188 (2001). URL http://dx.doi.org/10.1038/35056572.

[116] Rock, K. L., York, I. A., Saric, T. & Goldberg, A. L. Protein degradation and the generation of MHC class I-presented peptides. *Advances in Immunology* **80**, 1–70 (2002). URL http://www.sciencedirect.com/science/article/pii/S0065277602800128.

[117] Kim, H. & Yin, J. Effects of RNA splicing and post-transcriptional regulation on HIV-1 growth: a quantitative and integrated perspective. *Syst Biol* **152**, 138–152 (2005).

[118] Arnoczy, G. S. *et al.* Massive CD8 T Cell Response to Primary HIV Infection in the Setting of Severe Clinical Presentation. *AIDS Res Hum Retroviruses* **28**, 789–792 (2011).

[119] Brennan, C. a. *et al.* Early HLA-B*57-restricted CD8+ T lymphocyte responses predict HIV-1 disease progression. *Journal of Virology* **86**, 10505–16 (2012).

[120] Poropatich, K. & Sullivan, D. J. Human immunodeficiency virus type 1 long-term non-progressors: the viral, genetic and immunological basis for disease non-progression. *Journal of General Virology* **92**, 247–268 (2011).

[121] Bailey, J. R. *et al.* Transmission of human immunodeficiency virus type 1 from a patient who developed AIDS to an elite suppressor. *Journal of Virology* **82**, 7395–7410 (2008).

[122] Genovese, L., Nebuloni, M. & Alfano, M. Cell-mediated immunity in elite controllers naturally controlling hiv viral load. *Frontiers in Immunology* **4** (2013).

[123] Goulder, P. J. R. & Watkins, D. I. Impact of MHC class I diversity on immune control of immunodeficiency virus replication. *Nature reviews. Immunology* **8**, 619–30 (2008).

[124] Bailey, J. R., Williams, T. M., Siliciano, R. F. & Blankson, J. N. Maintenance of viral suppression in HIV-1-infected HLA-B*57+ elite suppressors despite CTL escape mutations. *The Journal of Experimental Medicine* **203**, 1357–69 (2006).

[125] Kelleher, a. D. *et al.* Clustered mutations in HIV-1 gag are consistently required for escape from HLA-B27-restricted cytotoxic T lymphocyte responses. *The Journal of Experimental Medicine* **193**, 375–386 (2001).

[126] Wagner, R. *et al.* Molecular and Functional Analysis of a Conserved CTL Epitope in HIV-1 p24 Recognized from a Long-Term Nonprogressor: Constraints on Immune Escape Associated with Targeting a Sequence Essential for Viral Replication. *The Journal of Immunology* **162**, 3727–3734 (1999).

[127] Masemola, A. M. *et al.* Novel and promiscuous CTL epitopes in conserved regions of Gag targeted by individuals with early subtype C HIV type 1 infection from southern Africa. *Journal of Immunology* **173**, 4607–4617 (2004).

[128] Goulder, P. J. *et al.* Novel, cross-restricted, conserved, and immunodominant cytotoxic T lymphocyte epitopes in slow progressors in HIV type 1 infection. *AIDS research and human retroviruses* **12**, 1691–1698 (1996).

[129] Miura, T. *et al.* HLA-B57/B*5801 Human Immunodeficiency Virus Type 1 Elite Controllers Select for Rare Gag Variants Associated with Reduced Viral Replication Capacity and Strong Cytotoxic T-Lymphotye Recognition. *Journal of Virology* **83**, 2743–2755 (2009).

[130] Crawford, H. *et al.* Compensatory mutation partially restores fitness and delays reversion of escape mutation within the immunodominant HLA-B*5703-restricted Gag epitope in chronic human immunodeficiency virus type 1 infection. *Journal of Virology* **81**, 8346–51 (2007).

[131] Kemal, K. S. *et al.* Transition from long-term nonprogression to HIV-1 disease associated with escape from cellular immune control. *Journal of Acquir. Immune Defic. Syndr* **48**, 119–126 (2008).

[132] Salgado, M. *et al.* An Additive Effect of Protective Host Genetic Factors Correlates with HIV Nonprogression Status. *Journal of Acquir. Immune Defic. Syndr* **56**, 300–305 (2011).

[133] Streeck, H. *et al.* Recognition of a defined region within p24 gag by CD8+ T cells during primary human immunodeficiency virus type 1 infection in individuals expressing protective HLA class I alleles. *Journal of Virology* **81**, 7725–7731 (2007).

[134] Troyer, R. M. *et al.* Variable fitness impact of HIV-1 escape mutations to cytotoxic T lymphocyte (CTL) response. *PLoS Pathogens* **5** (2009).

[135] Briggs, J. a. G. *et al.* The stoichiometry of Gag protein in HIV-1. *Nature structural & molecular biology* **11**, 672–675 (2004).

[136] Reddy, B. & Yin, J. Quantitative intracellular kinetics of HIV type 1. *AIDS research and human retroviruses* **15**, 273–283 (1999).

[137] Sacha, J. B. *et al.* Gag-Specific CD8+ T Lymphocytes Recognize Infected Cells before AIDS-Virus Integration and Viral Protein Expression. *Journal of Immunology* **178**, 2746–2754 (2007).

[138] Tang, J. *et al.* Human leukocyte antigen variants B*44 and B*57 are consistently favorable during two distinct phases of primary HIV-1 infection in sub-Saharan Africans with several viral subtypes. *Journal of Virology* **85**, 8894–902 (2011).

[139] Peterson, T. A. *et al.* HLA class I associations with rates of HIV-1 seroconversion and disease progression in the Pumwani Sex Worker Cohort. *Tissue Antigens* **81**, 93–107 (2013).

[140] Sette, A. *et al.* The relationship between class I binding affinity and immunogenicity of potential cytotoxic T cell epitopes. *The Journal of Immunology* **153**, 5586–5592 (1994). URL `http://www.jimmunol.org/content/153/12/5586`.

[141] Borghans, J. A. M., Mølgaard, A., de Boer, R. J. & Kemir, C. HLA alleles associated with slow progression to AIDS truly prefer to present HIV-1 p24. *PLoS ONE* **2** (2007).

[142] Paul, S. *et al.* HLA Class I Alleles Are Associated with Peptide-Binding Repertoires of Different Size, Affinity, and Immunogenicity. *The Journal of Immunology* **191**, 5831–5839 (2013). URL `http://www.jimmunol.org/content/191/12/5831`.

[143] MacNamara, A., Kadolsky, U., Bangham, C. R. M. & Asquith, B. T-cell epitope prediction: Rescaling can mask biological variation between MHC molecules. *PLoS Computational Biology* **5** (2009).

[144] Starcich, B. R. *et al.* Identification and characterization of conserved and variable regions in the envelope gene of HTLV-III/LAV, the retrovirus of AIDS. *Cell* **45**, 637–648 (1986).

[145] Wang, Y. & Lai, L. Modeling the intracellular dynamics for Vif-APO mediated HIV-1 virus infection. *Chinese Science Bulletin* **55**, 2329–2340 (2010).

[146] Kim, H. & Yin, J. Robust Growth of Human Immunodeciency Virus Type-1 (HIV-1). *Biophysical Journal* **89**, 2210–2221 (2005).

[147] Reddy, B. & Yin, J. Quantitative Intracellular Kinetics of HIV Type 1. *AIDS Research and Human Retroviruses* **15**, 273–283 (2004).

[148] Chen, Y. L., Trono, D. & Camaur, D. The proteolytic cleavage of human immunodeficiency virus type 1 Nef does not correlate with its ability to stimulate virion infectivity. *Journal of Virology* **72**, 3178–84 (1998).

[149] Mahalingam, S. *et al.* Identification of residues in the N-terminal acidic domain of HIV-1 Vpr essential for virion incorporation. *Virology* **207**, 297–302 (1995).

[150] Hockett, R. D. *et al.* Constant mean viral copy number per infected cell in tissues regardless of high, low, or undetectable plasma HIV RNA. *The Journal of experimental medicine* **189**, 1545–54 (1999).

[151] Mohammadi, P. *et al.* 24 hours in the Life of HIV-1 in a T Cell Line. *PLOS Pathogens* **9**, e1003161 (2013).

[152] Bohan, C. A. *et al.* Analysis of Tat transactivation of human immunodeficiency virus transcription in vitro. *Gene expression* **2**, 391–407 (1992).

[153] Graeble, M. A., Churcher, M. J., Lowe, A. D., Gait, M. J. & Karn, J. Human Immunodeciency Virus Type 1 Transactivator Protein, Tat, Stimulates Transcriptional Read-Through of Distal Terminator Sequences In Vitro. *Proceedings of the National Academy of Sciences of the United States of America* **90**, 6184–6188 (1993).

[154] Laspia, M. F., Wendel, P. & Mathews, M. B. HIV-1 Tat Overcomes Ineffcient Transcriptional Elongation in Vitro. *Molecular Biology* **232**, 732–746 (1993).

[155] Blanchard, J. M., Weber, J., Darnell, J. E. & Jelinek, W. In vitro RNA-RNA splicing in adenovirus 2 mRNA formation. *Proceedings of the National Academy of Sciences* **75**, 5344–5348 (1978).

[156] Love, D. C., Sweitzer, T. D. & Hanover, J. A. Reconstitution of HIV-1 rev nuclear export: independent requirements for nuclear import and export. *Proc Natl Acad Sci U S A* **95**, 10608–10613 (1998).

[157] Alberts, B. *et al. Essential Cell Biology* (Garland Publishing Inc, New York, 2003).

[158] Efthymiadis, A., Briggs, L. J. & Jans, D. A. The HIV-1 tat nuclear localization sequence confers novel nuclear import properties. *Journal of Biological Chemistry* **273**, 1623–1628 (1998).

[159] Sturniolo, T. *et al.* Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices. *Nature biotechnology* **17**, 555–561 (1999).

[160] Tong, J. C. *et al.* Prediction of HLADQ3.2 beta Ligands: Evidence of Multiple Registers in Class II Binding Peptides. *Bioinformatics* **22**, 1232–1238 (2006).

[161] Liao, W. W. P. & Arthur, J. W. Predicting Peptide Binding Affnities to MHC Molecules Using a Modied Semi-Empirical Scoring Function. *PLos ONE* **6**, 1–8 (2011).

[162] Chang, S. T. *Multi-Scale Modeling Of Antigen Presentation With Applications To Tuberculosis*. Ph.D. thesis, University of Michigan (2007).

[163] Bordner, A. J. & Mittlemann, H. D. Prediction of the Binding Anities of Peptides to Class II MHC Using a Regularized Thermodynamic Model. *BMC Bioinformatics* **11**, 1471–2105 (2010).

[164] Gakamsky, D. M., Davis, D. M., Strominger, J. L. & Pecht, I. Assembly and dissociation of human leukocyte antigen (HLA)-A2 studied by real-time fluorescence resonance energy transfer. *Biochemistry* **39**, 11163–11169 (2000).

[165] H.M. Eisen X.H. Hou, C. S. K. W., Tanguturi, V. K. & Smith, C. Promiscuous Binding of Extracellular Peptides to Cell Surface Class I MHC Protein. *Proc Natl Acad Sci U S A* **109**, 4580–4585 (2012).

[166] Dalchau, N. *et al.* A peptide ltering relation quanties MHC class I peptide optimization. *PLOS Computational Biology* **7**, e1002144 (2011).

[167] Lazaro, E. *et al.* Variable HIV peptide stability in human cytosol is critical to epitope presentation and immune escape. *Journal of Clinical Investigation* **121**, 2480–2492 (2011).

[168] Reits, E. *et al.* Peptide Diffusion, Protection, and Degradation in Nuclear and Cytoplasmic Compartments before Antigen Presentation by MHC Class I. *Immunity* **18**, 97–108 (2003).

[169] Yewdell, J. W., Reits, E. & Neefjes, J. Making sense of mass destruction: quantitating MHC class I antigen presentation. *Nature reviews. Immunology* **3**, 952–961 (2003).

[170] Serban, R. & Hindmarsh, A. C. {CVODES}, the sensitivity-enabled ODE solver in SUNDIALS. In *Proceedings of the 5th International Conference on Multibody Systems In Nonlinear Dynamics and Control* (ASME, Long Beach, CA, 2005).

[171] Serban, R. & Hindmarsh, a. C. CVODES: the Sensitivity-Enabled ODE Solver in SUNDIALS. *ACM Transactions on Mathematical Software* **5**, 1–18 (2003).

[172] Shehu-Xhilaga, M., Crowe, S. M. & Mak, J. Maintenance of the Gag/Gag-Pol Ratio Is Important for Human Immunodeficiency Virus Type 1 RNA Dimerization and Viral Infectivity. *Journal of Virology* **75**, 1834–1841 (2001).

[173] Kim, H. & Yin, J. Robust growth of human immunodeficiency virus type 1 (HIV-1). *Biophysical journal* **89**, 2210–21 (2005).

[174] Addo, M. M. *et al.* The HIV-1 regulatory proteins Tat and Rev are frequently targeted by cytotoxic T lymphocytes derived from HIV-1-infected individuals. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 1781–1786 (2001).

[175] Tsomides, T. J. *et al.* Naturally processed viral peptides recognized by cytotoxic T lymphocytes on cells chronically infected by human immunodeficiency virus type 1. *The Journal of Experimental Medicine* **180**, 1283–93 (1994).

[176] Brumme, Z. L. *et al.* Marked Epitope - and Allele- Specic Dierences in Rates of Mutation in Human Immunodeciency Type 1 (HIV-1) Gag, Pol, and Nef Cytotoxic T-Lymphocyte Epitopes in Acute/Early HIV-1 Infection. *Journal of Virology* **82**, 9216–9227 (2008).

[177] Henn, M. R. *et al.* Whole genome deep sequencing of HIV-1 reveals the impact of early minor variants upon immune recognition during acute infection. *PLOS Pathogens* **8**, e1002529 (2012).

[178] Rizzuto, C. D. *et al.* A conserved HIV gp120 glycoprotein structure involved in chemokine receptor binding. *Science* **280**, 1949–1953 (1998).

[179] Kloverpris, H. N. *et al.* Early Antigen Presentation of Protective HIV-1 KF11Gag and KK10Gag Epitopes from Incoming Viral Particles Facilitates Rapid Recognition of Infected Cells by Specific CD8+ T Cells. *Journal of Virology* **87**, 2628–2638 (2013).

[180] Ellis, E. L. & Delbrück, M. The Growth of Bacteriophage. *The Journal of General Physiology* **22**, 365–84 (1939). PMC2237944.

[181] Pedersen, M. & Phillips, A. Towards programming languages for genetic engineering of living cells. *Journal of The Royal Society Interface* **6**, S437 LP – S450 (2009). URL http://rsif.royalsocietypublishing.org/content/6/Suppl{_}4/S437.abstract.

[182] Nagorsen, D., Scheibenbogen, C., Marincola, F. M., Letsch, A. & Keilholz, U. Natural T Cell Immunity against Cancer. *Clinical Cancer Research* **9**, 4296 LP – 4303 (2003). URL http://clincancerres.aacrjournals.org/content/9/12/4296.abstract.

[183] Lee, P. P. *et al.* Characterization of circulating T cells specific for tumor-associated antigens in melanoma patients. *Nat Med* **5**, 677–685 (1999). URL http://dx.doi.org/10.1038/9525.

[184] Princiotta, M. F. *et al.* Quantitating Protein Synthesis, Degradation, and Endogenous Antigen Processing. *Immunity* **18**, 343–354 (2003). URL http://www.sciencedirect.com/science/article/pii/S1074761303000517.

[185] Schubert, U. *et al.* Rapid degradation of a large fraction of newly synthesized proteins by proteasomes. *Nature* **404**, 770–774 (2000). URL http://dx.doi.org/10.1038/35008096http://www.nature.com/nature/journal/v404/n6779/suppinfo/404770a0{_}S1.html.

[186] Yewdell, J. W. Amsterdamming DRiPs. *Molecular immunology* **55**, 110–112 (2013). URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3610827/.

[187] Lever, M. *et al.* Architecture of a minimal signaling pathway explains the T-cell response to a 1 million-fold variation in antigen affinity and dose. *Proceedings of the National Academy of Sciences* **113**, E6630–E6638 (2016). URL http://www.pnas.org/content/113/43/E6630.abstract.

[188] Eccleston, R. C., Wan, S., Dalchau, N. & Coveney, P. V. The role of multiscale protein dynamics in antigen presentation and T lymphocyte recognition. *Frontiers in Immunology* **8**, 797 (2017).